Bayesian motion estimation and segmentation

by

Yair Weiss

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1998

Author ....
Department of Brain and Cognitive Sciences
May 18, 1998

Certified by ....
Edward H. Adelson
Professor of Vision Science
Thesis Supervisor

Accepted by ....
Gerald E. Schneider
Chairman, Department Committee on Graduate Students

Bayesian motion estimation and segmentation

by

Yair Weiss

Submitted to the Department of Brain and Cognitive Sciences
on May 18, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Estimating motion in scenes containing multiple moving objects remains a difficult problem in computer vision yet is solved effortlessly by humans. In this thesis we present a computational investigation of this astonishing performance in human vision. The method we use throughout is to formulate a small number of assumptions and see the extent to which the optimal interpretation given these assumptions corresponds to the human percept.

For scenes containing a single motion we show that a wide range of previously published results are predicted by a Bayesian model that finds the most probable velocity field assuming that (1) images may be noisy and (2) velocity fields are likely to be slow and smooth. The predictions agree qualitatively, and are often in remarkable agreement quantitatively.

For scenes containing multiple motions we introduce the notion of "smoothness in layers". The scene is assumed to be composed of a small number of surfaces or layers, and the motion of each layer is assumed to be slow and smooth. We again formalize these assumptions in a Bayesian framework and use the statistical technique of mixture estimation to find the predicted percept. Again, we find a surprisingly wide range of previously published results that are predicted with these simple assumptions. We discuss the shortcomings of these assumptions and show how additional assumptions can be incorporated into the same framework.

Taken together, the first two parts of the thesis suggest that a seemingly complex set of illusions in human motion perception may arise from a single computational strategy that is optimal under reasonable assumptions. The third part of the thesis presents a computer vision algorithm that is based on the same assumptions. We compare the approach to recent developments in motion segmentation and illustrate its performance on real and synthetic image sequences.

Thesis Supervisor: Edward H. Adelson
Title: Professor of Vision Science

Acknowledgments

I thank my advisor, Ted Adelson, for his wisdom, support and encouragement. I hope that in my own career I will be able to follow the example that he sets — his endless curiosity about how the brain works, his brilliant insights into complicated phenomena and the fundamental honesty with which he conducts himself as a scientist. It has also been a pleasure to interact with Ted as a colleague and I will fondly remember our conversations on topics ranging from the theory of evolution to root beer.

Thanks also to the members of my thesis committee — Mike Jordan, Tommy Poggio and Jitendra Malik. Their suggestions and comments significantly improved the quality of this work.

I have benefited tremendously from discussions with students and colleagues over the five years I have been at MIT. Thanks to the many members of the Vision and Modeling group at the Media Lab and the Perceptual Science group at the Department of Brain and Cognitive Sciences. I learned a lot from conversations with John Wang, Josh Tenenbaum, Josh McDermott, Bart Anderson, Whitman Richards and many others.

I am sure that the friendships I formed here will continue well after I leave MIT. Special thanks go to members of the Eastgate Kibbutz including Luis, Adri, Yarden, Ori, Limor, Ariel, Yoel, Zohar, Amit and Rotem.

Last but not least I thank my parents, Yoram and Menucha. *Toda.*

Contents

Chapter 1

Introduction

The thesis is presented in terms of three self contained papers. However, the three papers are best read in sequence since they form part of a single line of study. In this introduction we make this line of inquiry explicit. We first describe the problem we wish to understand and the rationale for the method of investigation. We then summarize the main results of the three papers.

1.1 The Integration versus Segmentation Dilemma

In this thesis we seek to understand how the human visual system solves what we call "the integration versus segmentation dilemma". This dilemma arises from the conflicting demands of motion analysis in scenes containing multiple motions (Braddick, 1993). Due to the inherent ambiguity of local motion measurements, local computations do not gather enough information to obtain a correct estimate. Thus the system needs to *integrate* many local measurements. On the other hand, the fact that there are multiple motions means that global computations are likely to mix together measurements derived from different motions. Thus the system needs to *segment* the local measurements.

To illustrate these conflicting demands consider the simple scenes depicted in figures 1-1 through 1-3. Figure 1-1 shows the inherent ambiguity of local motion measurements. The well known "aperture problem" (Wallach, 1935; Horn and Schunck,

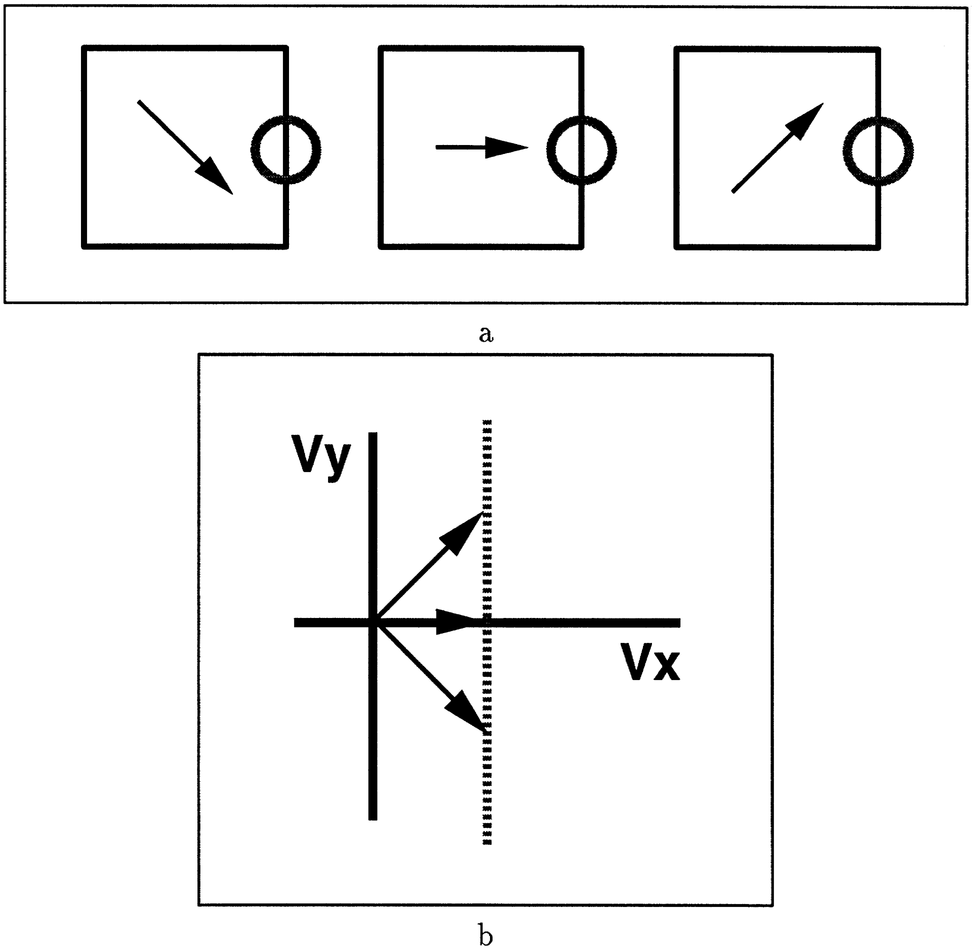


Figure 1-1: **a.** The “aperture problem” refers to the impossibility of determining the two dimensional motion of a signal containing a single orientation. Given a vertical edge, only the horizontal component of motion can be determined. **b.** The family of motions consistent with the motion of the edge can be depicted as a line in “velocity space”, where any velocity is represented as a vector from the origin whose length is proportional to speed and whose angle corresponds to direction of motion.

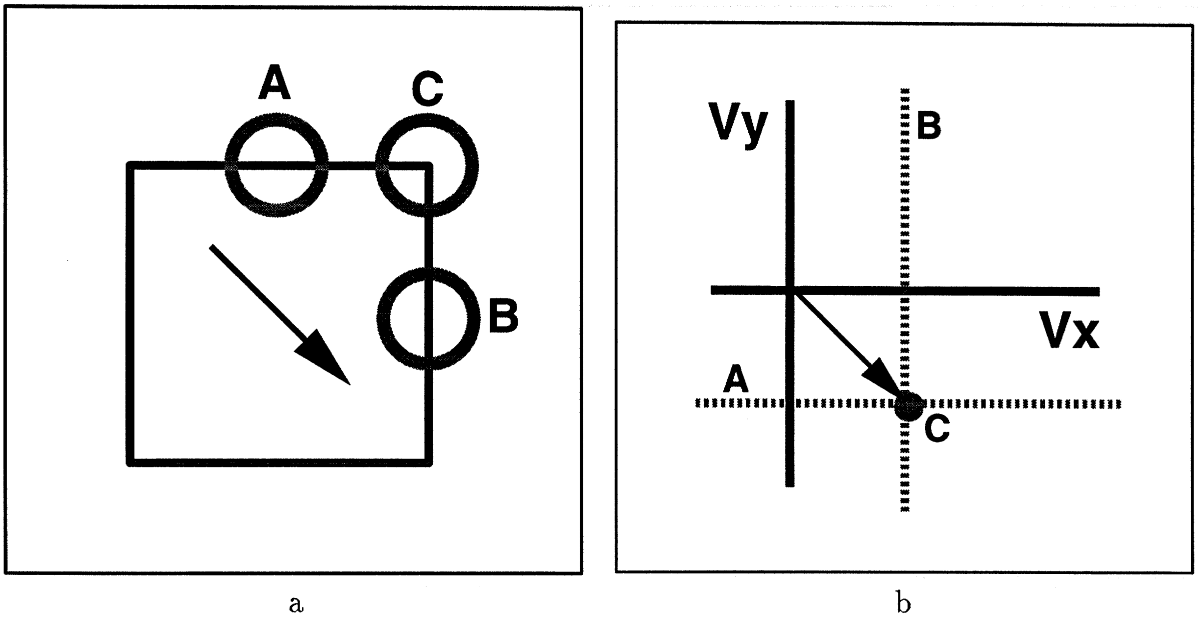


Figure 1-2: **a.** A single translating figure generates different constraints at different locations.**b.** When the constraints are plotted together they intersect at a single point yielding the physically correct velocity of the square. This is an example of how integrating multiple constraints can resolve the local ambiguity of motion measurements.

1981; Adelson and Movshon, 1982; Marr and Ullman, 1981; Fennema and Thompson, 1979) refers to the impossibility of determining the two dimensional motion when a signal only contains a single orientation. For example, a local analyzer that sees only the vertical edge of a square can only determine the horizontal component of the motion. Whether the square translates horizontally to the right, diagonally up and to the right, or diagonally down and to the right, the motion of the vertical edge will be the same. The family of motions consistent with the motion of the edge can be depicted as a line in "velocity space", where any velocity is represented as a vector from the origin whose length is proportional to speed and whose angle corresponds to direction of motion. Graphically, the aperture problem is equivalent to saying that the family of motions consistent with the information at an edge maps to a straight line in velocity space, rather than a single point.

This ambiguity may be reduced by combining information over space. Figure 1-2a shows velocity space representations of constraints from different image locations along the square. At an edge, the constraint is a line in velocity space with the same

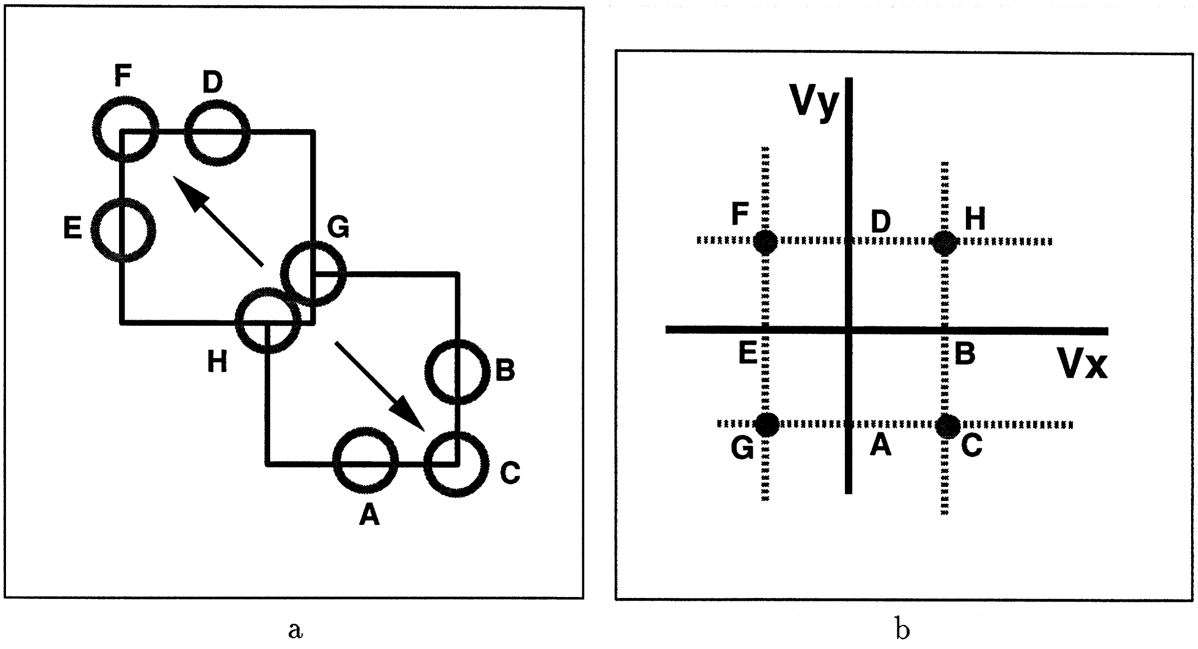


Figure 1-3: **a.** When the scene contains multiple objects, a simple velocity space construction is not sufficient to determine the motions. **b.** When the constraints from multiple locations are plotted together they intersect at four points. Two of these points correspond to the motion of a square, while the other two are spurious — they result from integrating together constraints that belong to different objects. This simple scene illustrates the need to simultaneously integrate and segment motion measurements.

orientation as the edge, while at a corner, the constraint is a point in velocity space — there is a single velocity consistent with the local data. Figure 1-2b shows all of these constraints plotted in single representation — they all intersect at a single point, and that intersection gives the physically correct velocity of the square.

If the visual system only needed to analyze motion in scenes containing a single object, it would only need to solve the integration problem, and not the segmentation problem. When the scene contains multiple objects, however, the situation is more complex. Figure 1-3 shows an example (after (Burt and Sperling, 1981)). Here there are two squares translating in different directions. As in the one-square case, measurements along the edge are ambiguous while measurements obtained at junctions are not. However, unlike the one-square case, the unambiguous measurements do not necessarily correspond to the correct motion of either square. Furthermore, when the constraints are plotted together they do not intersect at a single point. Rather four

Computational Theory	Representation and algorithm	Hardware Implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out.	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation.	How can the representation and algorithm be realized physically?

Figure 1-4: The three levels of analysis suggested by Marr and Poggio (reproduced from (Marr, 1982)). This thesis focuses on the computational theory level of explanation. We formulate a set of assumptions and constraints that may be used to analyze motion and compare the predicted percept to psychophysical data.

intersections are found. Two of these points correspond to the correct motions of the two squares. The other two points, however, do not correspond to any "true" motion in the scene. They are a result of mixing together constraints derived from different objects, of integrating together measurements that should be segmented.

As we show in this thesis, the simple two squares scene presents a problem for many computer vision motion analysis systems. Local motion analyzers cannot overcome the local ambiguity along the edges of the squares, while global approaches tend to mix together information belonging to different squares and predict an incorrect motion. Yet humans perceiving this scene have no trouble in indicating the motion at different points in the scene; the human visual system seems to have found a way to resolve the integration versus segmentation dilemma. The focus of this thesis is to understand this performance.

1.2 The computational approach to vision

> Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds'

> wings make sense. (Marr, 1982) p. 27

In order to understand how the human visual system resolves the "integration versus segmentation dilemma" we use the method of computational modeling. Marr and Poggio (Marr, 1982; Marr and Poggio, 1977) distinguished between three levels of understanding complex information-processing devices: the computational, the algorithmic and the implementation levels. Figure 1-4 reproduces Marr's summary of the three levels. The computational level is the most abstract; it describes the problem the system is trying to solve and the constraints it uses in order to solve it. The algorithmic level addresses questions of representations and the algorithm used to satisfy the constraints and assumptions of the system. Finally, the implementation level deals with the details of the hardware in which the algorithm is embodied.

Here we focus on the most abstract level, that of computational modeling. Rather than describing a particular biological implementation of a particular algorithm, we attempt to find the constraints and assumptions used by the visual system when estimating motion. The term "computational theory" is meant to emphasize the difference between this approach and a "verbal theory". A computational theory of vision requires more than a list of assumptions and constraints; rather these assumptions and constraints should be formulated in such a way that they can be computed for a given scene from the image data and therefore predict a percept for that scene.

In order to formalize our theory we employ a Bayesian inference framework (Knill and Richards, 1996). In the Bayesian formalism, the assumptions are formulated as probabilities and inference corresponds to finding the probability of hypotheses given observations. In the context of motion analysis, we express our assumptions in terms of *prior probabilities* — the probability of a motion hypothesis in the absence of any data and *likelihoods* — the probability of the image data given a motion hypothesis. We use the machinery of Bayesian inference to calculate the *posterior probability* — the probability of a motion hypothesis given the image data.

Finding the most probable motion hypothesis may require rather sophisticated computational algorithms. For example, in this thesis we use inversion of large matrices and an iterative algorithm known as Expectation-Maximization (Dempster et al., 1977). We do not, however, claim that the brain uses these algorithms. We use these algorithms here as tools in order to test our computational theory; the prediction of the theory is obtained by finding the most probable motion hypothesis for a given scene. We then compare this predicted motion to the human percept. If we find that the predicted motion matches the human percept, this supports our computational theory but says nothing about the algorithm or implementation the human visual system uses in order to arrive at the same percept.

1.3 Part 1 - A computational theory of motion integration

Rather than immediately taking on the full "integration versus segmentation" dilemma, the first part of the thesis deals only with the problem of integration. We address situations in which subjects are told that there is only a single moving surface in the scene and are asked to judge that motion. Thus there is no segmentation problem. The system still has to combine multiple local measurements to arrive at the motion of the surface. The first part asks how the human visual system performs this combination.

As mentioned in the introduction, given a stimulus with only a single orientation there is not enough information to find the two dimensional velocity vector. However, assuming the stimulus is translating in the image plane, two orientations should be sufficient to determine its velocity. Perhaps the simplest stimulus containing two orientations is the "plaid" stimulus in which two oriented gratings translate rigidly in the image plane (figure 1-5a). Due to the aperture problem, only the component of velocity normal to the orientation of the grating can be estimated, and hence each grating motion is consistent with an infinite number of possible velocities, a constraint line in velocity space (figure 1-5b). When each grating is viewed in isolation, subjects typically perceive the normal velocity (shown by arrows in figure 1-5b). Yet when the two gratings are presented simultaneously subjects often perceive them moving coherently and ascribe a single motion to the plaid pattern (Adelson and Movshon, 1982; Wallach, 1935).

Adelson and Movshon (1982) distinguished between three methods to estimate this "pattern motion" – Intersection of Constraints (IOC), Vector Average (VA) and blob tracking. Intersection of Constraints (IOC) finds the single translation vector that is consistent with the information at both gratings. Graphically, this can be thought of as finding the point in velocity space that lies at the intersection of both constraint lines (circle in figure 1-5b). Vector Average (VA) combines the two normal velocities by taking their average. Graphically this corresponds to finding the point in velocity space that lies halfway in between the two normal velocities (square in figure 1-5b). Blob tracking makes use of the motion of the intersections (Ferrera and Wilson, 1990; Mingolla et al., 1992) which contain unambiguous information indicating the pattern velocity. For plaid patterns blob tracking and IOC give identical predictions — they would both predict veridical perception.

The wealth of experimental results on the perception of motion in plaids reveals a surprisingly complex picture. Perceived pattern motion is sometimes veridical (consistent with IOC or feature tracking) and at other times significantly biased towards the VA direction. The degree of bias is influenced by factors including orientation of the gratings (Yo and Wilson, 1992; Bowns, 1996; Burke and Wenderoth, 1993), contrast (Stone et al., 1990), presentation time (Yo and Wilson, 1992) and foveal location (Yo and Wilson, 1992).

The insufficiency of any of the three mechanisms to explain the human percept may also be illustrated using the horizontally translating rhombus stimulus in figure 1-6. Although slightly more complicated than the plaid stimulus, the rhombus also contains only two orientations and may be used to test the three mechanisms suggested by Adelson and Movshon. A "narrow" rhombus whose corners are occluded (figure 1-6a) appears to move diagonally and is therefore consistent with VA but inconsistent with IOC or feature tracking. A "fat" rhombus whose corners are occluded (figure 1-6b) appears to move horizontally and is therefore consistent with IOC or feature tracking but not with VA. When the corners are visible (figure 1-6c) and the

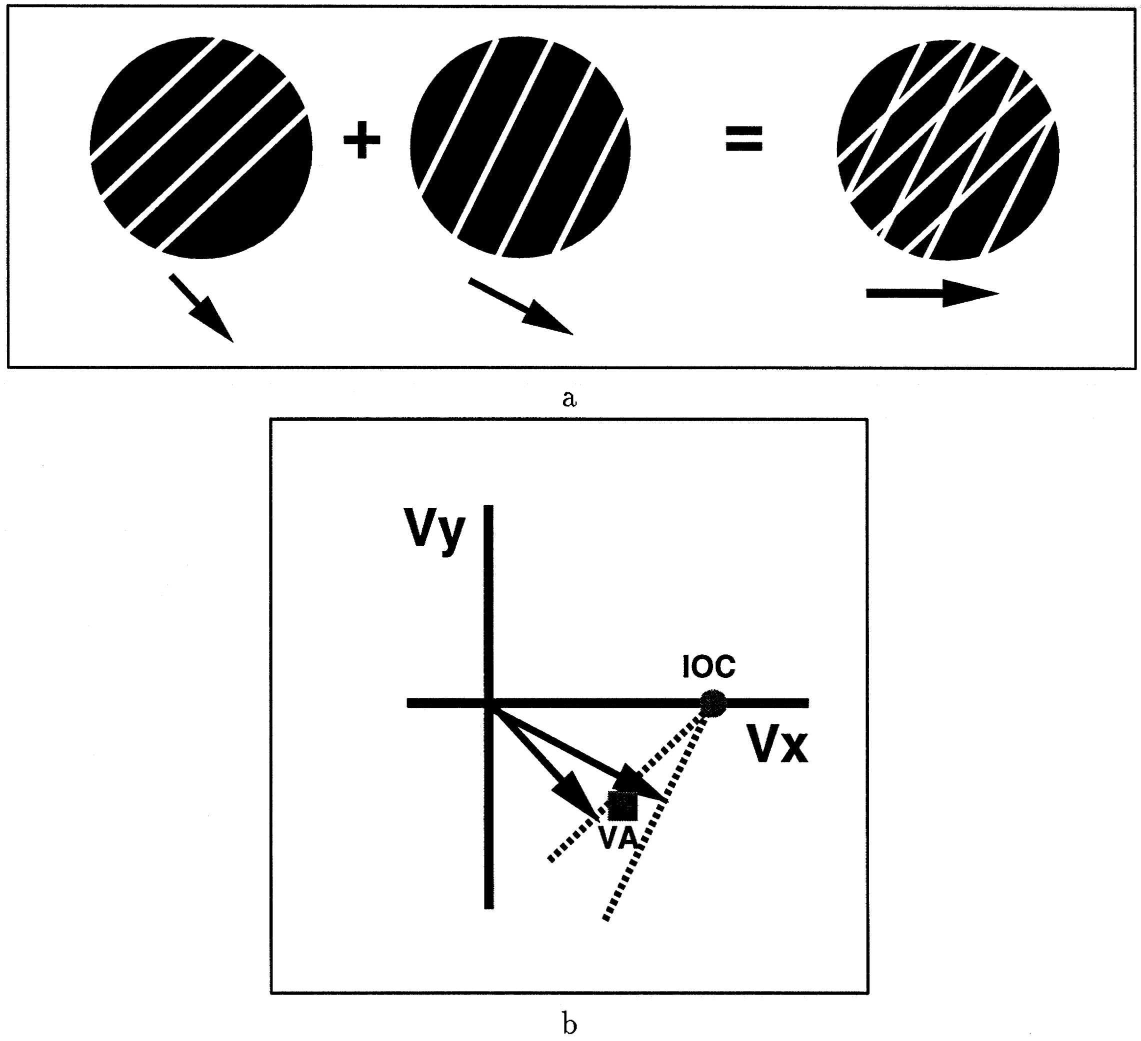


Figure 1-5: **a.** Two translating gratings form a "plaid" stimulus. Each grating by itself contains only a single orientation and its motion can not be estimated uniquely. However when both gratings are added, subjects often assign a single motion to the pattern (Adelson and Movshon, 1982; Wallach, 1935). For this reason, plaid stimuli are often used to study the method by which humans combine multiple motion measurements. **b.** Two velocity space constructions that could be used to estimate this "pattern motion" – Intersection of Constraints (IOC) and Vector Average (VA). Neither of these mechanisms are sufficient to to explain the experimental data on perceived direction of plaids. In the first part of the thesis we present a single computational mechanism that can explain the range of percepts.

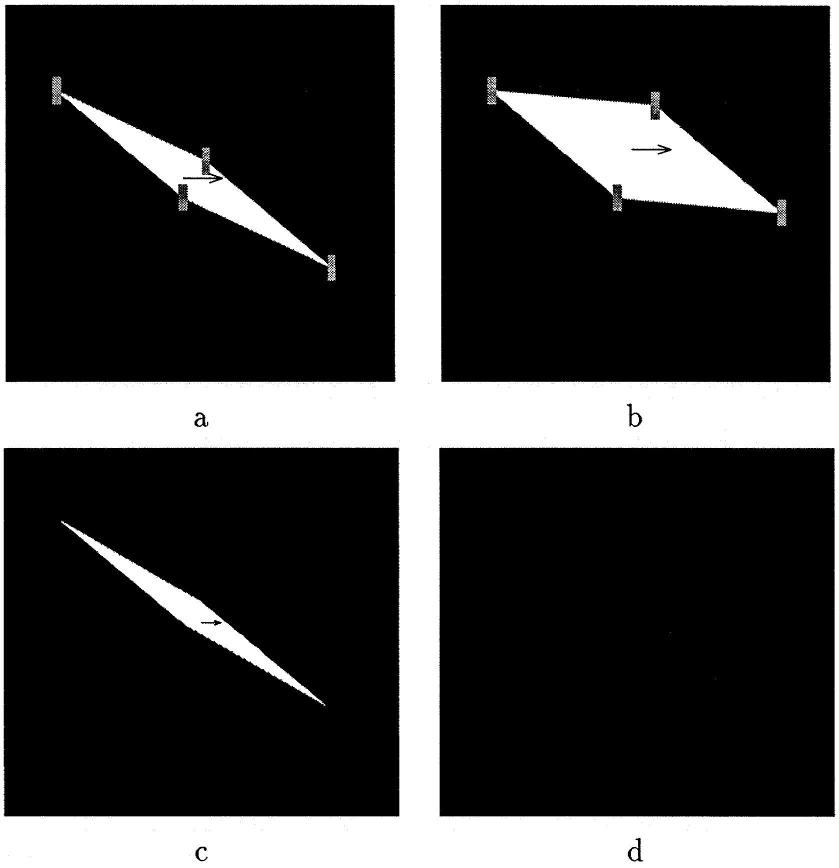


Figure 1-6: The insufficiency of either VA, IOC or feature tracking to explain the human percept. **a.** A "narrow" rhombus whose endpoints are occluded appears to move diagonally (consistent with VA). **b.** A "fat" rhombus whose endpoints are occluded appears to move horizontally (consistent with IOC or feature tracking). **c.** A high contrast "narrow" rhombus with visible endpoints appears to move horizontally (consistent with IOC or feature tracking). **d.** A low contrast "narrow" rhombus with visible endpoints appears to move diagonally (consistent with VA). In the first part of the thesis we show how these results are predicted by a single, simple model based on a few assumptions.

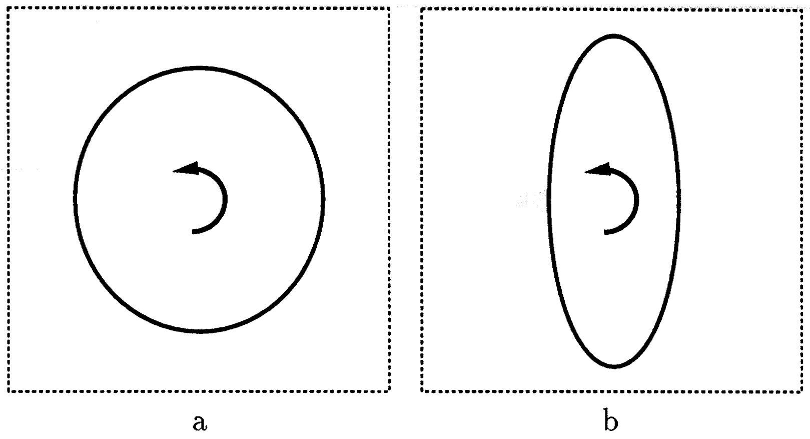


Figure 1-7: **a.** a "fat" ellipse rotating rigidly in the image plane appears to deform nonrigidly. **b.** a "narrow" ellipse rotating rigidly in the image plane appears to rotate rigidly. In the first part of the thesis we show that this phenomena is predicted by that the same assumptions that predict biases in perceived velocity of plaids and rhombuses.

contrast is high the rhombus appears to move horizontally, but when the contrast is low it appears to move diagonally.

Thus even for very simple stimuli, the question of which combination rule is used by human subjects is not easily answered. Conventional explanations involve multiple mechanisms, each invoked in order to explain a portion of the psychophysical results. The situation becomes even more complex when we consider non-translational motions. As an example, consider the perception of circles and derived figures in rotation (figure 1-7). When a "fat" ellipse , with aspect ratio close to unity, rotates in the image plane, it is perceived as deforming nonrigidly (Musatti, 1924; Wallach et al., 1956; Musatti, 1975). However, when a "narrow" ellipse, with aspect ratio far from unity, rotates in the image plane, the motion is perceived veridically (Wallach et al., 1956).

Obviously, none of the three mechanisms suggested for perception of plaids can be directly applied to explain this percept. These models estimate a single velocity vector rather than a spatially varying velocity field. An elegant explanation of the ellipse phenomena was offered by Hildreth (1983) using a very different style of model. She explained this and other motion "illusions" of smooth contours with a model that minimizes the variation of the perceived velocity field along the contour. She showed that for a rigid body with explicit features, her model will always give the physically "correct" motion field, but for smooth contours the estimate may be wrong. In the cases when the estimate was physically "wrong", it qualitatively agreed with human percepts of the same stimuli. Grzywacz and Yuille (1991) used a modified definition of smoothness to explain the misperception of smooth contours undergoing rigid translation (Nakayama and Silverman, 1988a; Nakayama and Silverman, 1988b). Poggio et al. (85) have shown that the smoothness assumption is useful in many aspects of computational vision.

In the first part of the thesis, we show that a single, simple model based on a small number of assumptions can account for a wide range of percepts. The model is essentially based on only two assumptions: (1) a likelihood term that assumes that image measurements may be noisy and (2) a prior term that favors *slow and smooth* velocity fields. We calculate the velocity field that maximizes the posterior probability and compare this prediction to the percept of human observers. In reviewing a long list of previously published phenomena, we find that the Bayesian estimator almost always predicts the psychophysical results. The predictions agree qualitatively, and are often in remarkable agreement quantitatively.

1.4 Part 2 - A computational theory of motion estimation and segmentation

In the second part of the thesis, we address the larger question — simultaneous integration and segmentation of motion constraints. The first part showed that the assumption of slow and smooth velocity fields can account for the human percept in a wide range of scenes containing a single motion. However, if the scene contains multiple objects, the "slow and smooth" assumption predicts percepts that are nothing like the human percept. Figure 1-8 shows the predicted velocity for the two squares scene – while humans perceive two rigid bodies, the "slow and smooth" assumption predicts a single, elastically deforming body. In the second part of the thesis

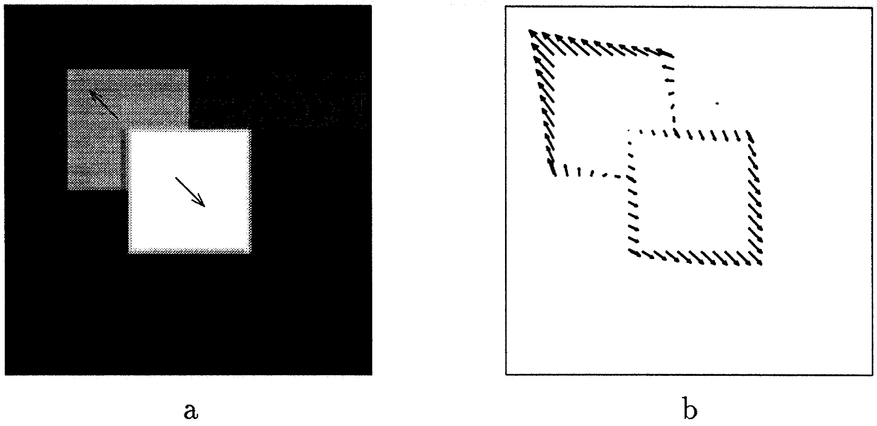


Figure 1-8: **a.** Two squares translate in the image in different directions. **b.** The output of a standard smoothness algorithm on this sequence. The algorithm tries to simultaneously fit the motion of both surfaces and recovers a single elastic deformation that is not at all like the human percept.

we present an extension of the "slow and smooth" assumption to scenes containing multiple motions.

The failure of global smoothness algorithms in scenes containing multiple motions such as figure 1-8 is well known and several ways of fixing the smoothness assumption have been proposed. Hildreth (1983) proposed a model whereby smoothness is only assumed along contours. Her algorithm found the velocity field of least variation along the zero crossings of the image. To illustrate her assumption consider the two squares scene discussed in figure 1-8. Hildreth's algorithm would first extract contours from this scene and then combine measurements along the contour. Thus assuming that the first step correctly extracted two contours, one for the boundary of each square, her algorithm would only assume smoothness in the motion of each square. It would not assume any relationship between the motions of the two squares. Thus for this stimulus it would predict two rigid motions, consistent with human perception.

Although Hildreth's assumption of smoothness along contours does solve some of the problems associated with smoothness models, there is reason to believe it is not exactly the assumption used by the human visual system. As pointed out by Grzywacz and Yuille (1991) the Hildreth assumption would predict no influence between features that are off the contour and the perceived motion of the contour.

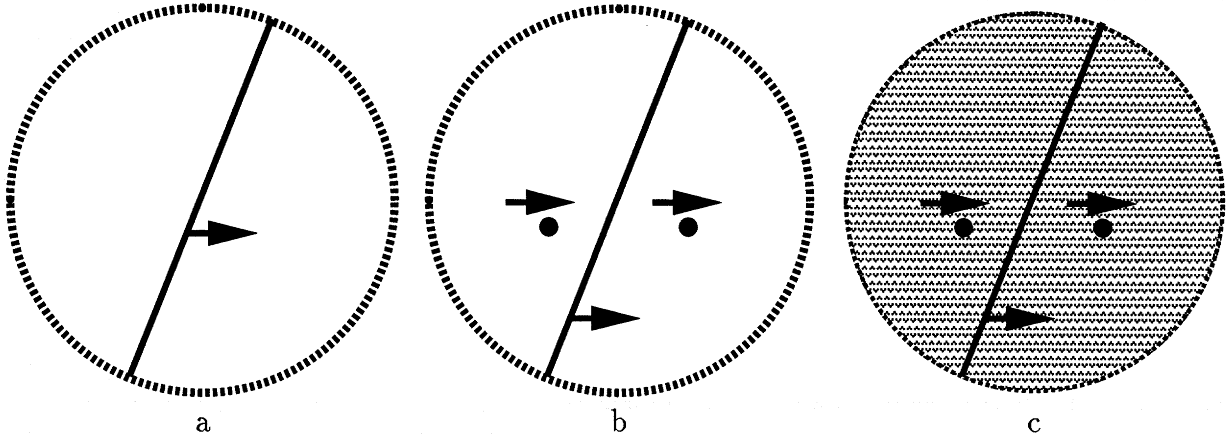


Figure 1-9: **a.** A horizontally translating diagonal line whose endpoints are invisible is consistent with an infinite family of motions. Typically, under these conditions, the normal velocity is chosen and the line appears to translate diagonally. (Wallach 35) **b.** When two horizontally translating dots are added to the display the line appears to move in the direction of the dots (Wallach 35, Rubin and Hochstein 93). This is inconsistent with a model that only combines information along contours (e.g. Hildreth 83). **c.** The effect persists when the display is placed on a static texture background. This is inconsistent with an algorithm that assumes "smoothness with discontinuities" (e.g. Terzopoulos 86). The discontinuities formed between the dots and the background would inhibit any interactions between the dots and the line.

This is inconsistent with experimental results that show a strong influence of features in such displays (e.g.(Nakayama and Silverman, 1988a; Shiffrar et al., 1995; Weiss and Adelson, 1995; Rubin and Hochstein, 1993)). Figure 1-9 shows an example dating back to Wallach (1935). A line whose endpoints are invisible appears to move in the normal direction, but when a small number of dots translating horizontally are added to the display they tend to "capture" the line, and the line appears to move horizontally. The fact that the dot influences the line when it is not part of the line's contour is inconsistent with Hildreth's model or any other model that assumes smoothness only along contours.

Rather than restricting the smoothness assumption to contours, other approaches assume a smooth two dimensional velocity field with possible discontinuities. In these models, e.g. (Terzopoulos, 1986; Hutchinson et al., 1988; Horn, 1986), nearby points are assumed to have similar velocities, but if the velocities are too dissimilar the assumption is abandoned and a discontinuity is assumed there instead. An advantage of these models over the standard smoothness models is that when the location of the discontinuity is estimated correctly, there is no smoothing across boundaries. This avoids many of the oversmoothing problems associated with global smoothness algorithms.

Despite these successes, there exist scenes in which the discontinuities approach predicts a motion that is very different from the human percept. Essentially, it predicts no interaction between two locations if there is a motion discontinuity between them. Figure 1-9c shows a simple example in which the line and the dot translate horizontally over a static background.The dissimilarity between the motions of the dots and the background texture would give rise to a discontinuity as would the dissimilarity between the line and the texture. Yet human perceiving this scene report no difference between the percept with and without the static texture. The dots and the line appear to be in front of the texture and are perceived as a single surface. Thus while piecewise smoothness may be a reasonable assumption to make in many contexts, it does not appear to be sufficient for modeling human motion perception.

As these simple demonstrations show, the visual system does not appear to assume global smoothness over the image, nor does it assume smoothness only along contours, nor does it assume smoothness with discontinuities. In the second part of the thesis, we propose a formulation that we call "smoothness in layers". We assume the scene includes a small number of surfaces or layers (Wang and Adelson, 1994) and that motion varies smoothly within a given layer. To illustrate this assumption consider figure 1-10. Global smoothness would assume that motion varies smoothly over the entire image, while smoothness in layers assumes that one velocity field will vary smoothly over the front surface and a second velocity field will vary smoothly over the back surface. There is *no* assumption of smoothness between two layers only within layers. This distinction is illustrated in a $1D$ example in figure 1-11

Unfortunately, the input to the visual system is not a description in terms of surfaces or layers. Thus if we wish to account for human motion perception by assuming smoothness in layers, we need to also account for the formation of a layered description from spatiotemporal data. In the second part of the thesis we present a

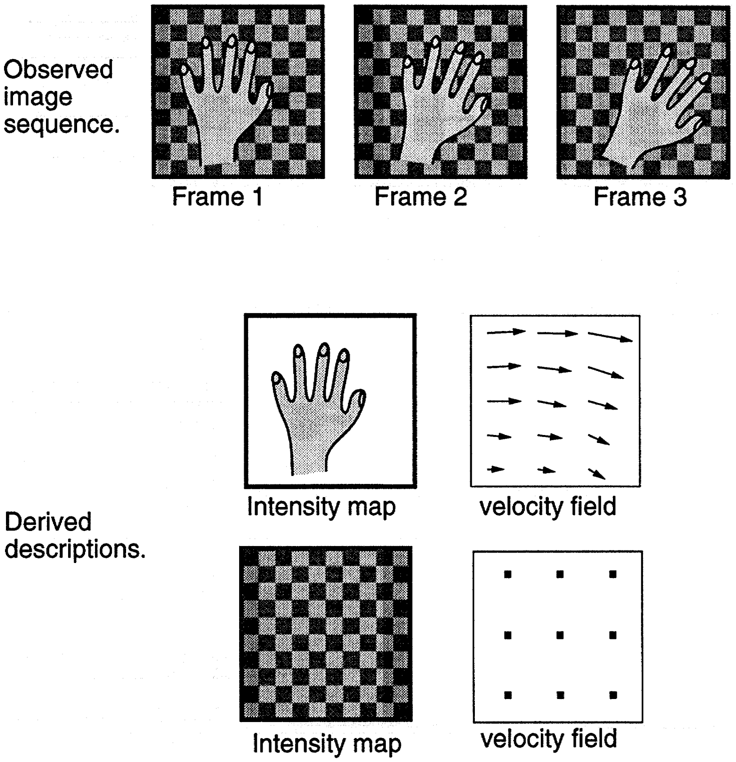


Figure 1-10: Layered decomposition of image sequences (adapted from (Wang and Adelson, 1994)). In a layered description, an image sequence is decomposed into a small number of occluding layers or surfaces, and each layer has a corresponding motion field. In this paper we propose that human motion perception assumes the motion field of each layer is smooth, but does not assume smoothness between motion fields of different layers.

computational model that receives as input a gray level image sequence and calculates (1) the number of layers (2) the assignment of pixels to layers and (3) the velocity field of each layer.

The model uses the statistical framework of mixture estimation to find the most probable interpretation of a scene. It is based on three assumptions: (1) a likelihood term identical to the one used in the first part that assumes image measurements may be noisy (2) a prior term that favors slow and smooth velocity fields *within a layer* and (3) a preference for a small number of layers. In order to validate these assumptions, we compare the most probable interpretation under these assumption to human percepts in previously published stimuli.

As an example of the type of data we would like to account for, consider figure 1-

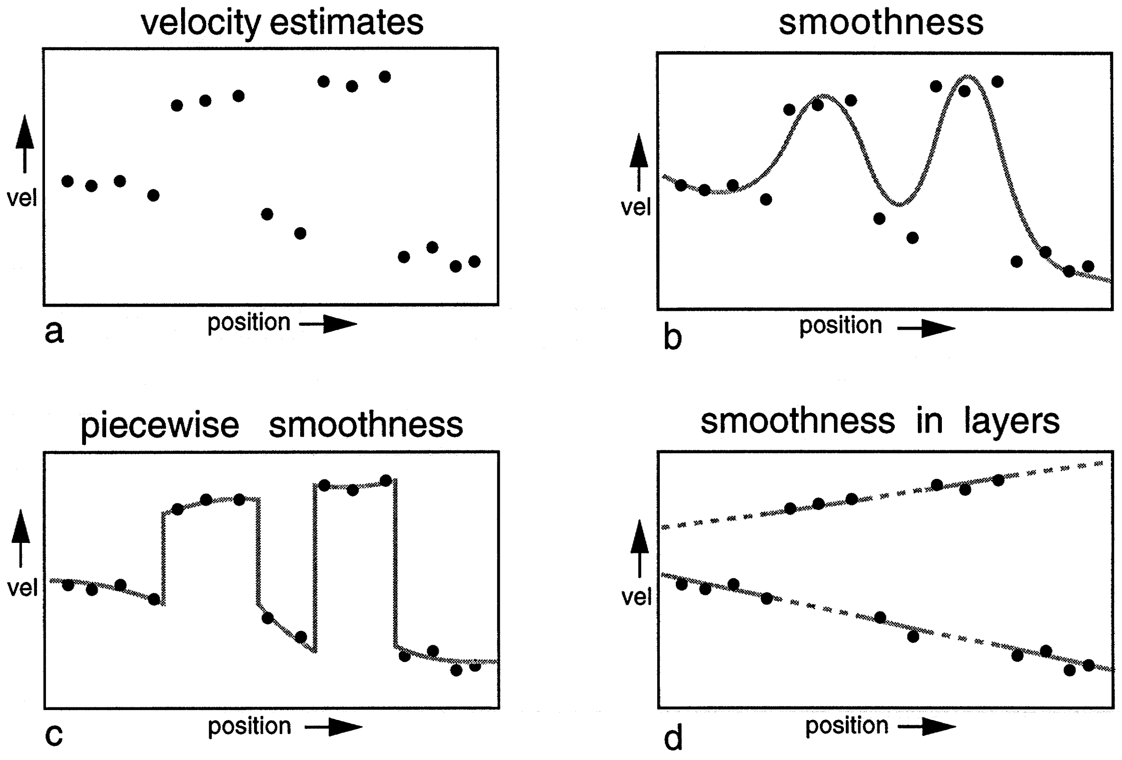


Figure 1-11: An illustration of the smoothness in layers assumption in $1D$ (adapted from (Wang and Adelson, 1994)). **a.** Hypothetical velocity estimates as a function of position. Such data would typically arise from two surfaces in depth. **b.** Global smoothness assumption applied to this data. The measurements from the two surfaces are mixed together rather than segmented. **c.** Piecewise smoothness. Information is not propagated across discontinuities. The resulting estimate is rather noisy. **d.** Smoothness in Layers. Two smooth velocity functions are found, one for each surface.

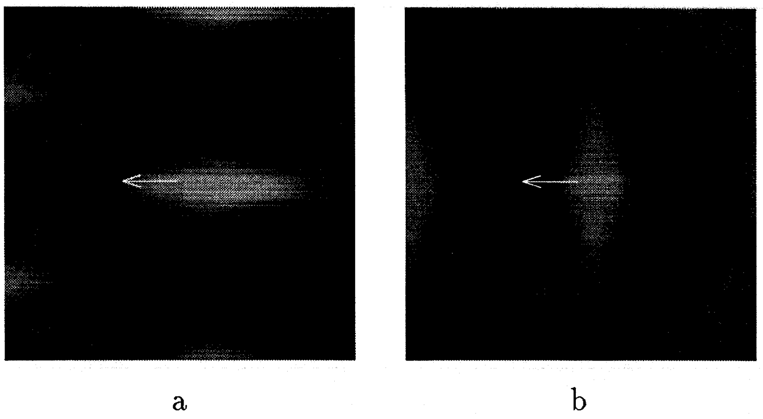


Figure 1-12: Adelson and Movshon (1982) found that the tendency of plaids to cohere depended on the difference between the principal direction of the two gratings. Thus the plaid in **a** tends to cohere less than the plaid in **b**. In the second part of the thesis we show that this tendency is predicted by the assumption of "smoothness in layers".

12. When the plaid on the left is shown to human subjects, they tend to see two motions — the plaid does not cohere but each grating is seen as moving in its normal direction. However when the plaid on the right is presented, subjects tend to see a single, coherent pattern translating in the horizontal direction. The tendency of plaids to cohere or appear transparent has been widely studied and has shown to be influenced by speed, period, orientation and contrast (Adelson and Movshon, 1982; Kim and Wilson, 1993; Farid and Simoncelli, 1994). As we show in the second part of the thesis, these tendencies are predicted from the three assumptions outlined above — there is no need for stimulus specific heuristics.

Although plaids present the most widely studied stimuli in which humans were asked to judge whether one or two motions were present, the "smoothness in layers" assumption is by no means restricted to plaid stimuli. Figure 1-13 shows the split herringbone illusion (Adelson and Movshon, 1983). Two sets of diagonal lines translate vertically in opposite directions. At high contrast (Adelson and Movshon, 1983) the percept consists of two groups, one moving up and the other moving down. However, if the stimulus is blurred, viewed peripherally or at low contrast, one perceives a single coherent motion to the right. As we show in the second part, this "illusory" motion at low contrast is actually the most probable interpretation given the assumption of

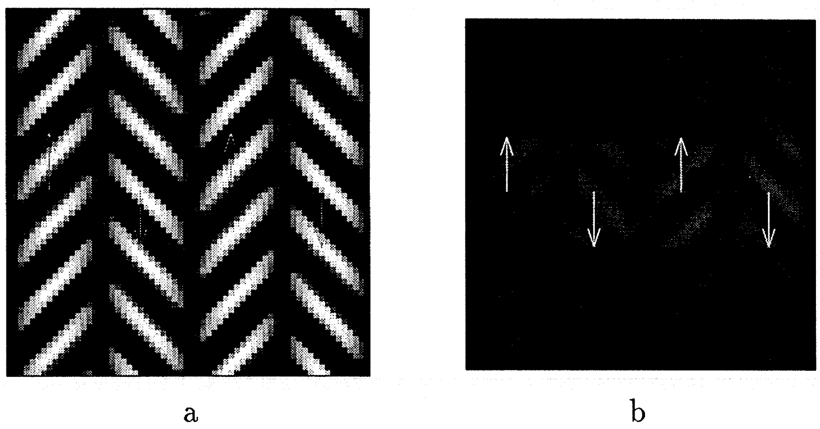


Figure 1-13: The split herringbone illusion (Adelson and Movshon, 1983). Two sets of diagonal lines translate vertically in opposite directions. At high contrast (Adelson and Movshon, 1983) the percept consists of two groups, one moving up and the other moving down. However, if the stimulus is blurred, viewed peripherally or at low contrast, one perceives a single coherent motion to the right. In the second part of the thesis, we show that this "illusion" is the most probable percept given the "smoothness in layers" assumption.

smoothness in layers.

Since we estimate a smooth velocity field for every layer, the smoothness in layers assumption can be applied to scenes in which the objects are undergoing nontranslational motions. Figure 1-14 shows an example with the ellipse stimulus discussed earlier. When a "fat" ellipse rotates rigidly in the image plane it is perceived as deforming nonrigidly (Wallach et al., 1956). When four rotating dots are added to the display, the ellipse and the dots are perceived as moving together and the ellipse is perceived as rigid (Weiss and Adelson, 1995). The effect of the four satellites persists when a large number of vertically translating dots is added to the display (Weiss and Adelson, 1995). In this case, humans perceive two groups — the ellipse and the four dots are perceived as moving together but the vertically translating dots are perceived as being in a separate layer. As we show in the second part of the thesis, this tendency is also predicted by the smoothness in layers assumption.

Despite the success of the smoothness in layers assumption in accounting for a wide range of stimuli, there exist stimuli for which these three assumptions are not sufficient. The second part of the thesis also discusses these shortcomings. There are

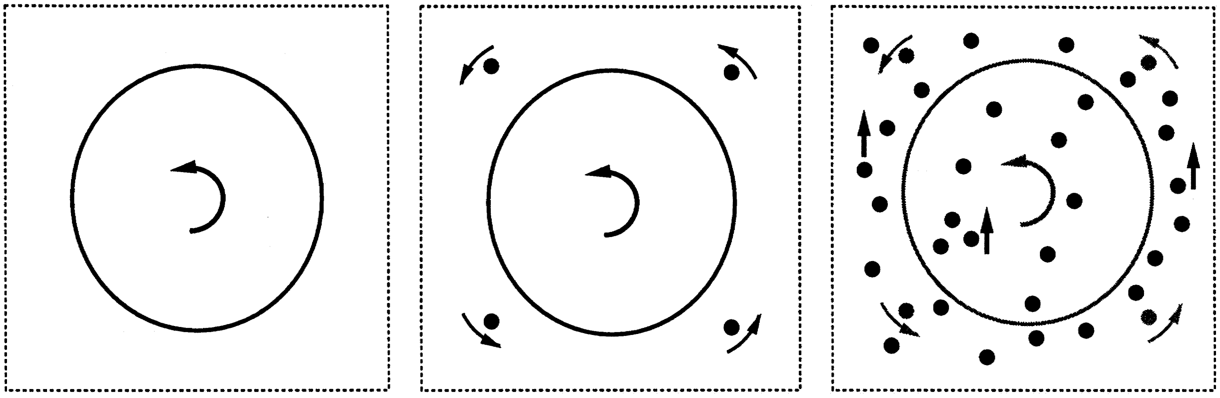


Figure 1-14: When a "fat" ellipse rotates rigidly in the image plane it is perceived as deforming nonrigidly (Wallach et al., 1956). When four rotating dots are added to the display, the ellipse is perceived as rigid (Weiss and Adelson, 1995). The effect of the satellites persists when a large number of vertically translating dots is added to the display (Weiss and Adelson, 1995). In the second part of the thesis we show that this is also predicted by the "smoothness in layers" assumption.

many non-motion cues that influence the tendency of humans to segment or integrate motion measurements. For example in the case of plaids, cues such as stereo depth, the luminance of the intersections and the relative spatial frequencies cause the plaid to appear more transparent (Adelson and Movshon, 1982; Stoner et al., 1990; Bressan et al., 1993). These cues increase the tendency to see the plaid as two transparent gratings even in a single static frame, and the three assumptions discussed above know nothing about this static analysis. We discuss how to augment these assumptions so they can incorporate additional cues and show preliminary results with the more sophisticated set of assumptions.

1.5 Part 3 - Integration and Segmentation in machine vision

In section 1.1 we discussed the inherent ambiguity of local motion measurements. For the synthetic ideal images we discussed the ambiguity only occurs at locations containing a single orientation. Locations such as corners, where multiple orientations exist locally, are unambiguous. However, in the presence of noise, all local measurements

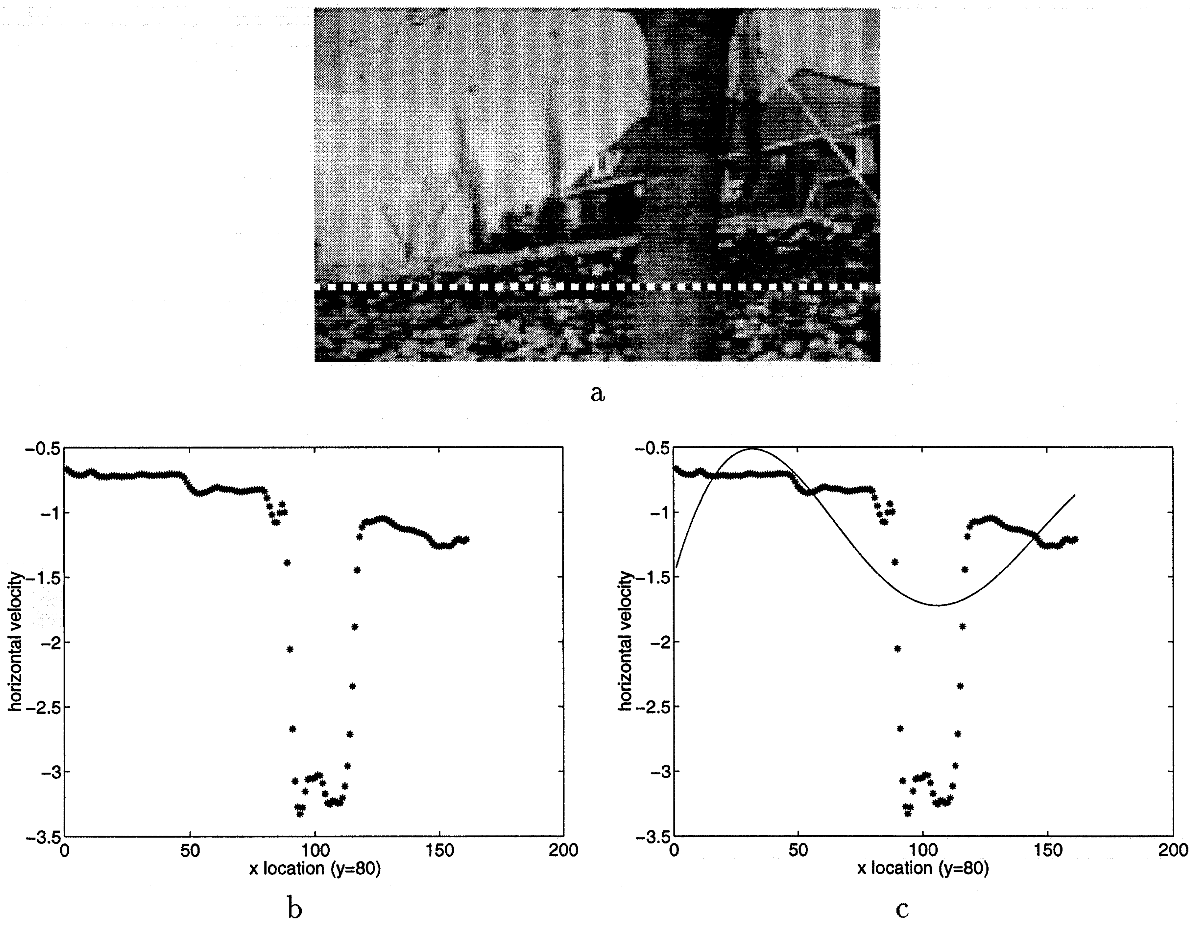


Figure 1-15: The integration versus segmentation dilemma in natural images. **a.** A single frame from the MPEG flower garden sequence. The sequence was shot by a camera placed on a driving car, and the image motion is related to distance from the camera. Thus the tree, which is closest to the camera moves fastest. **b.** The output of state-of-the art local motion analyzer on this scene (Bergen et al., 1992). We show the horizontal estimated velocity at a cross section of the image. Note that even though all locations are textured and hence contain multiple orientations, the estimated local flow is still quite noisy. **c.** The output of a global smoothness algorithm on this sequence. The tree is predicted to move much slower than it really does because its motion is combined with that of the flowers. The algorithm is integrating constraints that should be segmented.

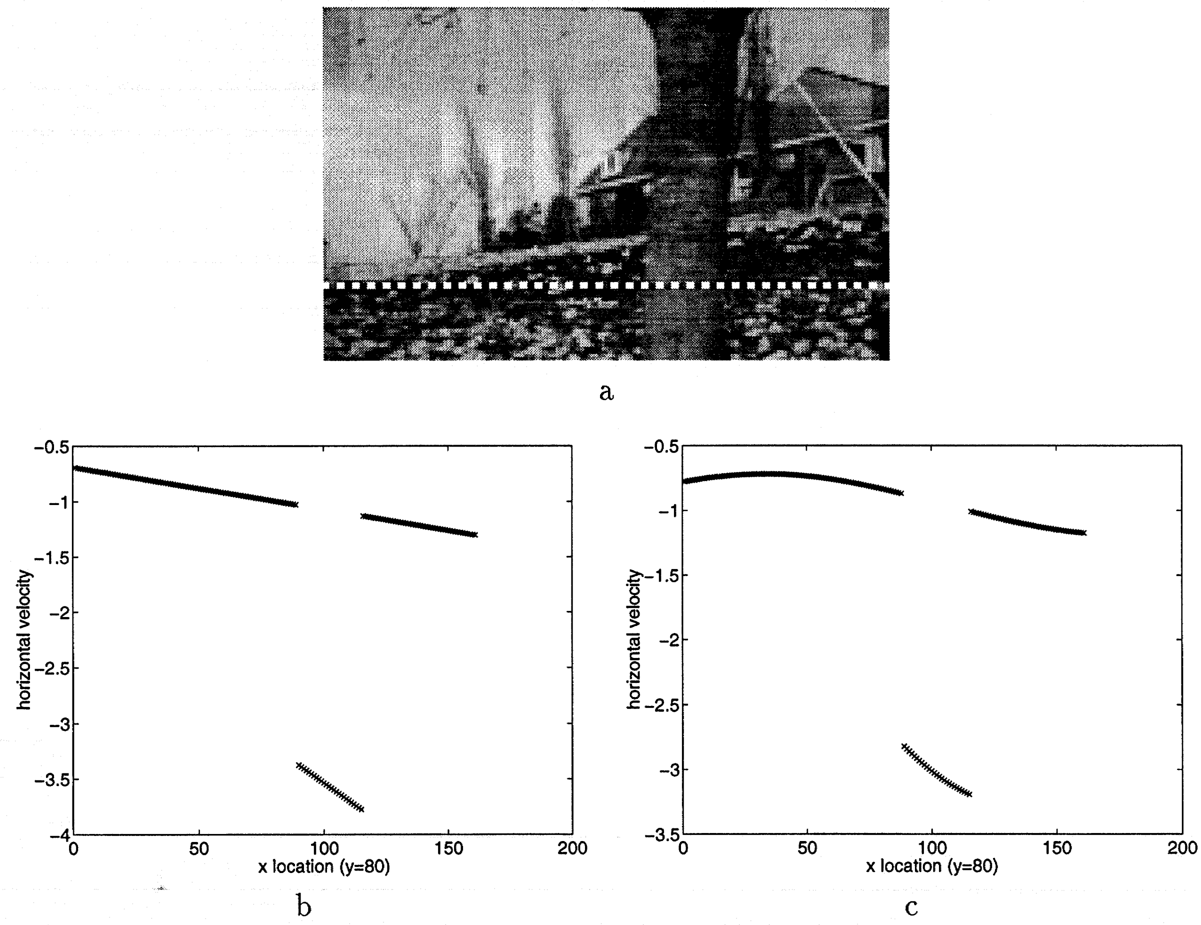


Figure 1-16: The output of motion segmentation algorithms on the flower garden sequence. **b.** A cross section through the horizontal flow field predicted by the Wang and Adelson (1984) algorithm. Although the algorithm segments the tree from the flowers, the motion of each layer is flat as if it were made of cardboard. **c.** A cross section through the horizontal flow field predicted by the smoothness in layers algorithm. Note that the algorithm more accurately captures the curved shape of the tree and the flower beds.

are ambiguous — even if the location contains multiple orientations the extraction of the correct constraints is limited by the noise in the image. Thus in a noisy world, the distinction between "corners" and "lines" is a slightly artificial dichotomy — all locations have some ambiguity. Rather than "ambiguous" versus "unambiguous" locations, in a noisy world the degree of ambiguity of local measurements may take on a continuum of values.

Figure 1-15 shows an example of the ambiguity of local measurements in real world scenes. The sequence was shot by a camera placed on a driving car, and the image motion is related to distance from the camera. Thus the tree, which is closest to the camera moves fastest. Figure 1-15a shows the output of state-of-the art local motion analyzer on this scene (Bergen et al., 1992). We show the horizontal estimated velocity at a cross section of the image. The pixels corresponding to the tree move fastest and the background pixels corresponding to the flower bed move with spatially varying slower speed. Note that even though all locations are textured and hence contain multiple orientations, the estimated local flow is still quite noisy. At the border of the tree and the flower garden the local analysis gives an intermediate velocity that is quite different from the true image motion, and along the flower bed the local estimate varies noisily from location to location in a way that does not reflect the true depth of the scene. Figure 1-15c shows the output of a global smoothness algorithm on this sequence. Again we show the only the horizontal estimated flow along a cross section. Now the estimate is highly smooth but quite wrong. Since the algorithm assumes a smoothly varying velocity field, the tree is predicted to move much slower than it really does. Its motion is influenced by the motion of the slowly moving flowers. This is precisely the integration versus segmentation dilemma — deriving reliable estimates requires integrating information from multiple locations while segmenting information derived from different motions.

The problems with global smoothness approaches are well known and have prompted a recent trend in computer vision towards approaches that fit multiple, global motion models to the image data. (Darrell and Pentland, 1991; Jepson and Black, 1993; Irani and Peleg, 1992; Hsu et al., 1994; Ayer and Sawhney, 1995; Wang and Adelson, 1994).

While differing in implementation, these algorithms share the goal of deriving from the image data a representation consisting of (1) a small number of global motion models and (2) a segmentation map that indicates which pixels are assigned to which model.

In order to segment images based on common motion, most existing algorithms assume that the motion of each model is described by a low dimensional parameterization. The two most popular choices are a six parameter affine model (Wang and Adelson, 1994; Weiss and Adelson, 1996) or an eight parameter projective model (Ayer and Sawhney, 1995; Irani and Peleg, 1992). Both of these parameterizations correspond to the rigid motion of a plane: the affine model assumes orthographic projection while the projective model assumes a perspective projection.

Despite the success of existing algorithms in segmenting image sequences, the assumption that motion segments correspond to rigid planar patches is obviously restrictive. Non-planar surfaces, or objects undergoing non-rigid motion cannot be grouped. Even when the segmentation is correct, the restriction to planar motions means that the estimated motion for each segment may be wrong. Figure 1-16b shows the estimated motion from a segmentation algorithm that assumes planar motions (Wang and Adelson, 1994; Ayer and Sawhney, 1995; Weiss and Adelson, 1996). Figure 1-16b shows a cross section from the estimated flow — at each location we plot the horizontal velocity of the segment to which that location belongs. Note that unlike the global smoothness approach, the tree does not "pull along" portions of the flower bed. The constraints from the tree and the bed are segmented rather than integrated. However, the motion of the tree and the flower bed are both approximated by planar motions — as if they were painted on flat sheets of cardboard.

In the third part of the thesis we show that the "smoothness in layers" assumption can be used to segment such scenes. We present an algorithm that segments such scenes and estimates a smooth motion field for each layer or segment rather than a low-dimensional parametric flow field. We combine the ideas of smoothness and segmentation in a mixture model framework, and this leads to an efficient Expectation-Maximization (EM) algorithm. An additional advantage of the mixture estimation framework is that additional cues for segmentation can be incorporated in a natural fashion. Figure 1-16c shows the output of our algorithm on the flower garden sequence. The segmentation is similar to the one obtained using planar motions but the estimated motion is rather different. Unlike the planar models, the velocity fields have enough degrees of freedom in order to capture the curved nature of the tree and the flower bed. Unlike the global smoothness algorithm, the algorithm avoids mixing together constraints derived from different objects. These results suggest that the same assumptions used in modeling the psychophysical result may also be useful in improving the performance of computer vision systems.

Chapter 2

Slow and Smooth: a Bayesian theory for the combination of local motion signals in human vision

Abstract

In order to estimate the motion of an object, the visual system needs to combine multiple local measurements, each of which carries some degree of ambiguity. We present a model of motion perception whereby measurements from different image regions are combined according to a Bayesian estimator — the estimated motion maximizes the posterior probability assuming a prior favoring slow and smooth velocities. In reviewing a large number of previously published phenomena we find that the Bayesian estimator predicts a wide range of psychophysical results. This suggests that the seemingly complex set of illusions arise from a single computational strategy that is optimal under reasonable assumptions.

2.1 Introduction

Estimating motion in scenes containing multiple, complex motions remains a difficult problem for computer vision systems, yet is performed effortlessly by human observers. Motion analysis in such scenes imposes conflicting demands on the design of a vision system (Braddick, 1993). The inherent ambiguity of local motion signals means that local computations cannot provide enough information to obtain a correct

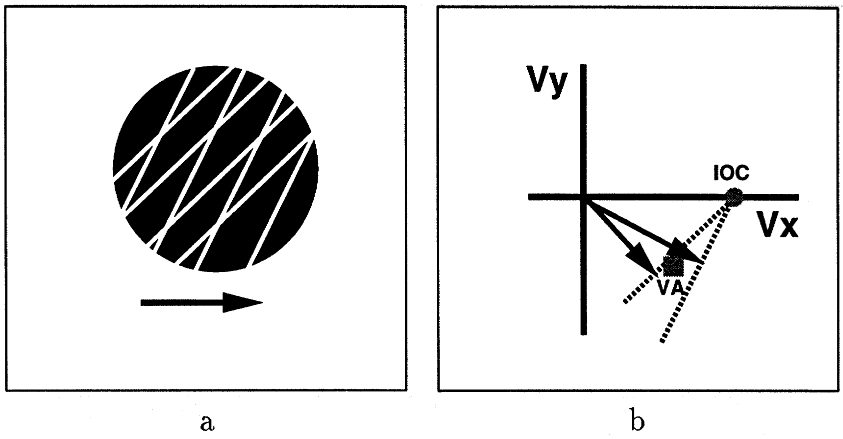


Figure 2-1: **a.** Two gratings translating in the image plane give a "plaid" pattern. **b.** Due to the aperture problem, the measurements for a single grating are consistent with a family of motions all lying on a constraint line in velocity space. Intersection of Constraints (IOC) finds the single velocity consistent with both sources of information. Vector Averaging (VA) takes the average of the two normal velocities. Experimental evidence for both types of combination rules has been found.

estimate. Thus the system must *integrate* many local measurements. On the other hand, the fact that there are multiple motions means that global computations are likely to mix together measurements derived from different motions. Thus the system also must *segment* the local measurements.

In this paper we are concerned with the first part of the problem, the integration of multiple constraints. Even if we know the scene contains only a single object, estimating that motion is nontrivial. This difficulty arises from the ambiguity of individual velocity measurements which may give only a partial constraint on the unknown motion (Wallach, 1935) , i.e. the "aperture problem", (Horn and Schunck, 1981; Adelson and Movshon, 1982; Marr and Ullman, 1981). To solve this problem, most models assume a two stage scheme whereby local readings are first computed, and then integrated in a second stage to produce velocity estimates. Psychophysical (Adelson and Movshon, 1982; Movshon et al., 1986; Welch, 1989) and neurophysiological (Movshon et al., 1986; Rodman and Albright, 1989) findings are consistent with such a model.

The nature of the integration scheme used in the second stage remains, however, controversial. This is true even for the simple, widely studied "plaid" stimulus in which two oriented gratings translate rigidly in the image plane (figure 2-1a). Due to the aperture problem, only the component of velocity normal to the orientation of the grating can be estimated, and hence each grating motion is consistent with an infinite number of possible velocities, a constraint line in velocity space (figure 2-1b). When each grating is viewed in isolation, subjects typically perceive the normal velocity (shown by arrows in figure 2-1b). Yet when the two gratings are presented simultaneously subjects often perceive them moving coherently and ascribe a single motion to the plaid pattern (Adelson and Movshon, 1982; Wallach, 1935).

Adelson and Movshon (1982) distinguished between three methods to estimate this "pattern motion" – Intersection of Constraints (IOC), Vector Average (VA) and blob tracking. Intersection of Constraints (IOC) finds the single translation vector that is consistent with the information at both gratings. Graphically, this can be thought of as finding the point in velocity space that lies at the intersection of both constraint lines (circle in figure 2-1b). Vector Average (VA) combines the two normal velocities by taking their average. Graphically this corresponds to finding the point in velocity space that lies halfway in between the two normal velocities (square in figure 2-1b). Blob tracking makes use of the motion of the intersections (Ferrera and Wilson, 1990; Mingolla et al., 1992) which contain unambiguous information indicating the pattern velocity. For plaid patterns blob tracking and IOC give identical predictions — they would both predict veridical perception.

The wealth of experimental results on the perception of motion in plaids reveals a surprisingly complex picture. Perceived pattern motion is sometimes veridical (consistent with IOC or feature tracking) and at other times significantly biased towards the VA direction. The degree of bias is influenced by factors including orientation of the gratings (Yo and Wilson, 1992; Bowns, 1996; Burke and Wenderoth, 1993), contrast (Stone et al., 1990), presentation time (Yo and Wilson, 1992) and foveal location (Yo and Wilson, 1992).

Thus even for the restricted case of plaid stimuli, neither of the three models suggested above can by themselves explain the range of percepts. Instead, one needs to assume that human motion perception is based on at least two separate mechanisms — a "2D motion" mechanism that estimates veridical motion and a crude "1D motion" mechanism that is at times biased away from the veridical motion. Many investigators have proposed that two separate motion mechanisms exist and that these are later combined (Rubin and Hochstein, 1993; Lorenceau et al., 1992; Mingolla et al., 1992; Alais et al., 1994).

As an example of a two mechanism explanation, consider the Wilson et al. (92) model of perceived direction of sine wave plaids. The perceived motion is assumed to be the average of two motion estimates one obtained by a "Fourier" pathway and the other by a "non-Fourier" pathway. The "Fourier" pathway calculates the normal motions of the two components while the "non-Fourier" pathway calculates motion energy on a squared and filtered version of the pattern.

Both pathways use vector average to calculate their motion estimates, but the inclusion of the "non-Fourier" pathway causes the estimate to be more veridical. Wilson et al. have shown that their model may predict biased or veridical estimates of direction depending on the parameters of the stimulus. The change in model prediction with stimulus parameters arises from the fact that the two mechanisms operate in separate regimes. Thus since plaids move in the vector average at short durations and not at long durations, it was assumed that the "non-Fourier" mechanism is delayed relative to the "Fourier" pathway. Since plaids move more veridically in the fovea than in the periphery, the model non-Fourier responses were divided by two in the periphery.

The danger of such an explanation is that practically any psychophysical result on perceived direction can be accommodated - by assuming that the "2D" mechanism operates when the motion is veridical, and does not operate whenever the motion is biased. For example, Alais et al (1994) favor a 2D "blob tracking" explanation for perceived direction of plaids. The fact that some plaids exhibit large biases in perceived direction while others do not is attributed to the fact that some plaids contain "optimal blobs" while others contain "suboptimal blobs" (Alais et al., 1994). Although the data may require these types of post-hoc explanations, we would prefer a more principled explanation in terms of a single mechanism.

Evidence that the complex set of experimental results on plaids may indeed be

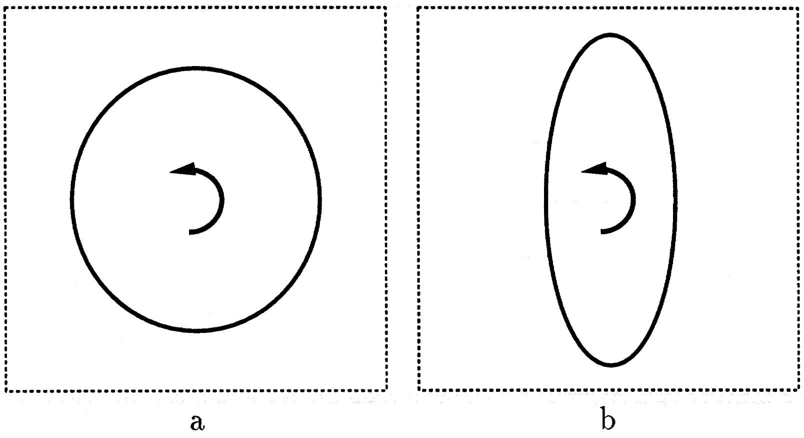


Figure 2-2: **a.** a “fat” ellipse rotating rigidly in the image plane appears to deform nonrigidly. **b.** a “narrow” ellipse rotating rigidly in the image plane appears to rotate rigidly.

explained using a single principled mechanism comes from the work of Heeger and Simoncelli (Heeger and Simoncelli, 1991; Simoncelli and Heeger, 1992; Simoncelli, 1993; Simoncelli and Heeger, 1998). Their model consisted of a bank of spatiotemporal filters, whose outputs were pooled to form velocity tuned units. The population of velocity units represented an optimal Bayesian estimate of the local velocity, assuming a prior probability favoring slow speeds. Their model worked directly on the raw image data and could be used to calculate the local velocity for any image sequence. In general, their model predicted a velocity close to the veridical velocity of the stimulus, but under certain conditions (e.g. low contrast, small angular separation) predicted velocities that were biased towards the vector average. They showed that these conditions for biased perception were consistent with data from human observers.

The controversy over the integration scheme used to estimate the translation of plaids may obscure the fact that they are hardly representative of the range of motions the visual system needs to analyze. A model of integration of local constraints in human vision should also account for perception of more complex motions than rigid 2D translation in the image plane. As an example, consider the perception of circles and derived figures in rotation (figure 2-2). When a “fat” ellipse , with aspect ratio close to unity, rotates in the image plane, it is perceived as deforming nonrigidly (Musatti, 1924; Wallach et al., 1956; Musatti, 1975). However, when a "narrow" ellipse, with aspect ratio far from unity, rotates in the image plane, the motion is perceived veridically (Wallach et al., 1956).

Unfortunately, the models surveyed above for the perception of plaids can not be directly applied to explain this percept. These models estimate a single velocity vector rather than a spatially varying velocity field. An elegant explanation was offered by Hildreth (1983) using a very different style of model. She explained this and other motion "illusions" of smooth contours with a model that minimizes the variation of the perceived velocity field along the contour. She showed that for a rigid body with explicit features, her model will always give the physically "correct" motion field, but for smooth contours the estimate may be wrong. In the cases when the estimate was physically "wrong", it qualitatively agreed with human percepts of the same stimuli. Grzywacz and Yuille (1991) used a modified definition of smoothness to explain the misperception of smooth contours undergoing rigid translation (Nakayama and Silverman, 1988a; Nakayama and Silverman, 1988b).

Thus the question of how the visual system integrates multiple local motion constraints has not a single answer in the existing literature but rather a multitude of answers. Each of the models proposed can successfully explain a subset of the rich experimental data.

In this paper we propose a single Bayesian model for motion integration and show that it can account for a wide range of percepts. We show that seemingly unconnected phenomena in human vision – from bias towards vector average in plaids to perceived nonrigidity in ellipses may arise from an optimal Bayesian estimation strategy in human vision.

2.2 Intuition — Bayesian motion perception

In order to obtain intuition about how Bayesian motion perception works, this section describes the construction of an overly simplified Bayesian motion estimator. As we discuss at the end of this section, this restricted model *can not* account for the range

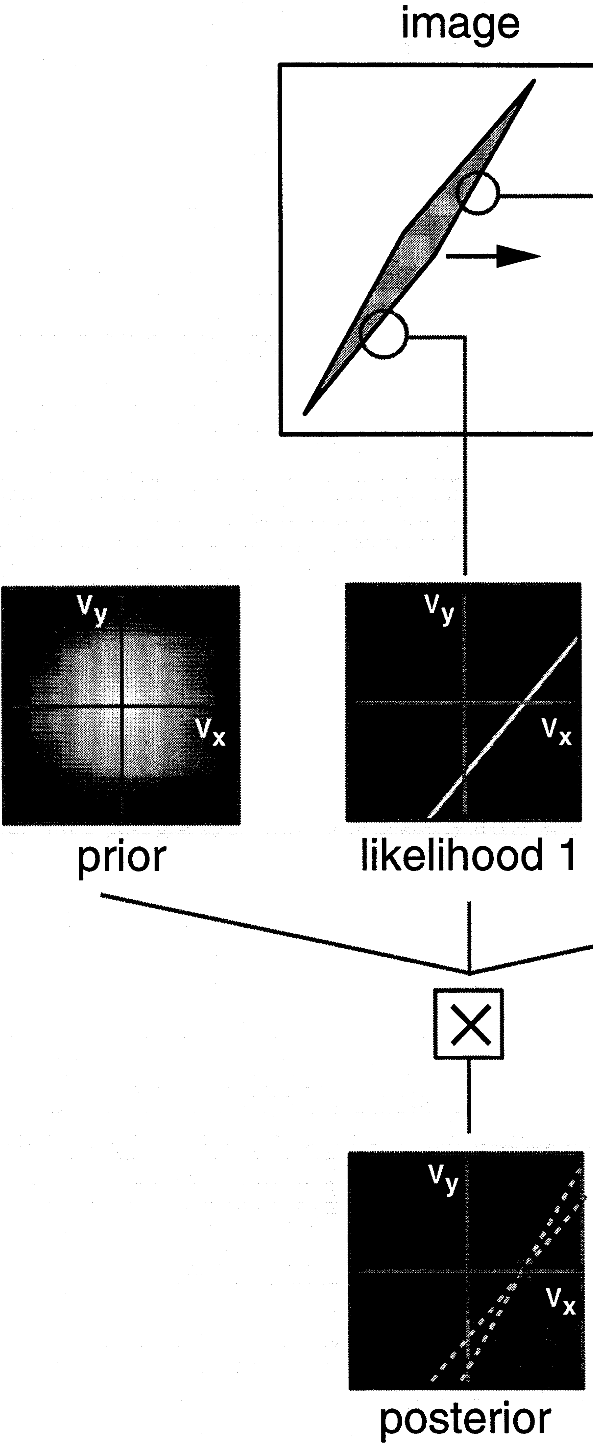


Figure 2-3: A restricted Bayesian estimator for velocity. The algorithm receives local likelihoods from various image locations and calculates the posterior probability in velocity space. This estimator is too simplistic to account for the range of phenomena we are intersted in explaining but serves to give intuition about how Bayesian motion estimation works. Here the likelihoods are zero everywhere except on the constraint line and the MAP estimate is the IOC solution.

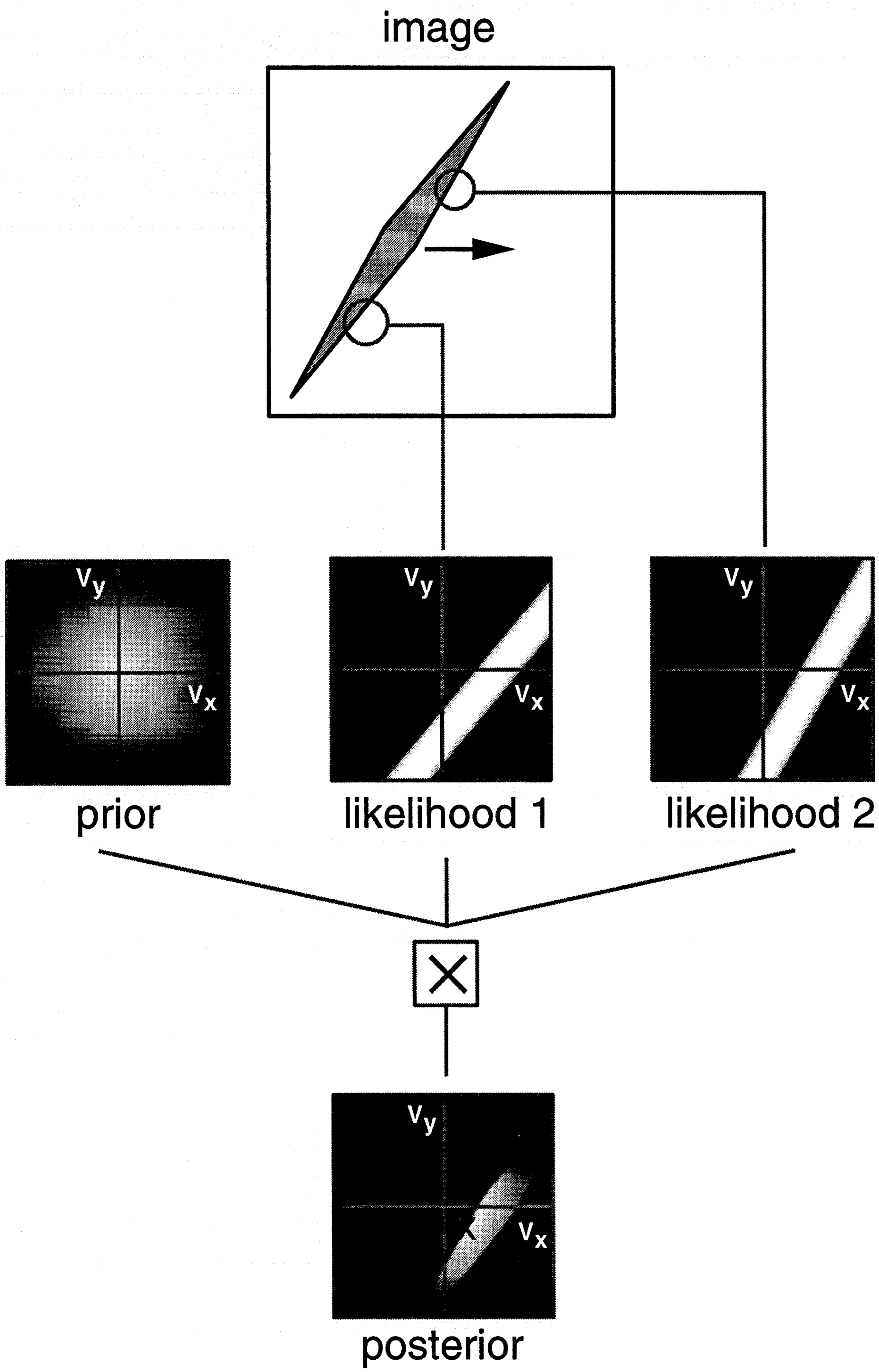


Figure 2-4: A restricted Bayesian estimator for velocity. The algorithm receives local likelihoods from various image locations and calculates the posterior probability in velocity space. This estimator is too simplistic to account for the range of phenomena we are interested in explaining but serves to give intuition about how Bayesian motion estimation works. Here the likelihoods are zero everywhere except at distance ϵ from the constraint line and the MAP estimate is the normal velocity with minimal speed.

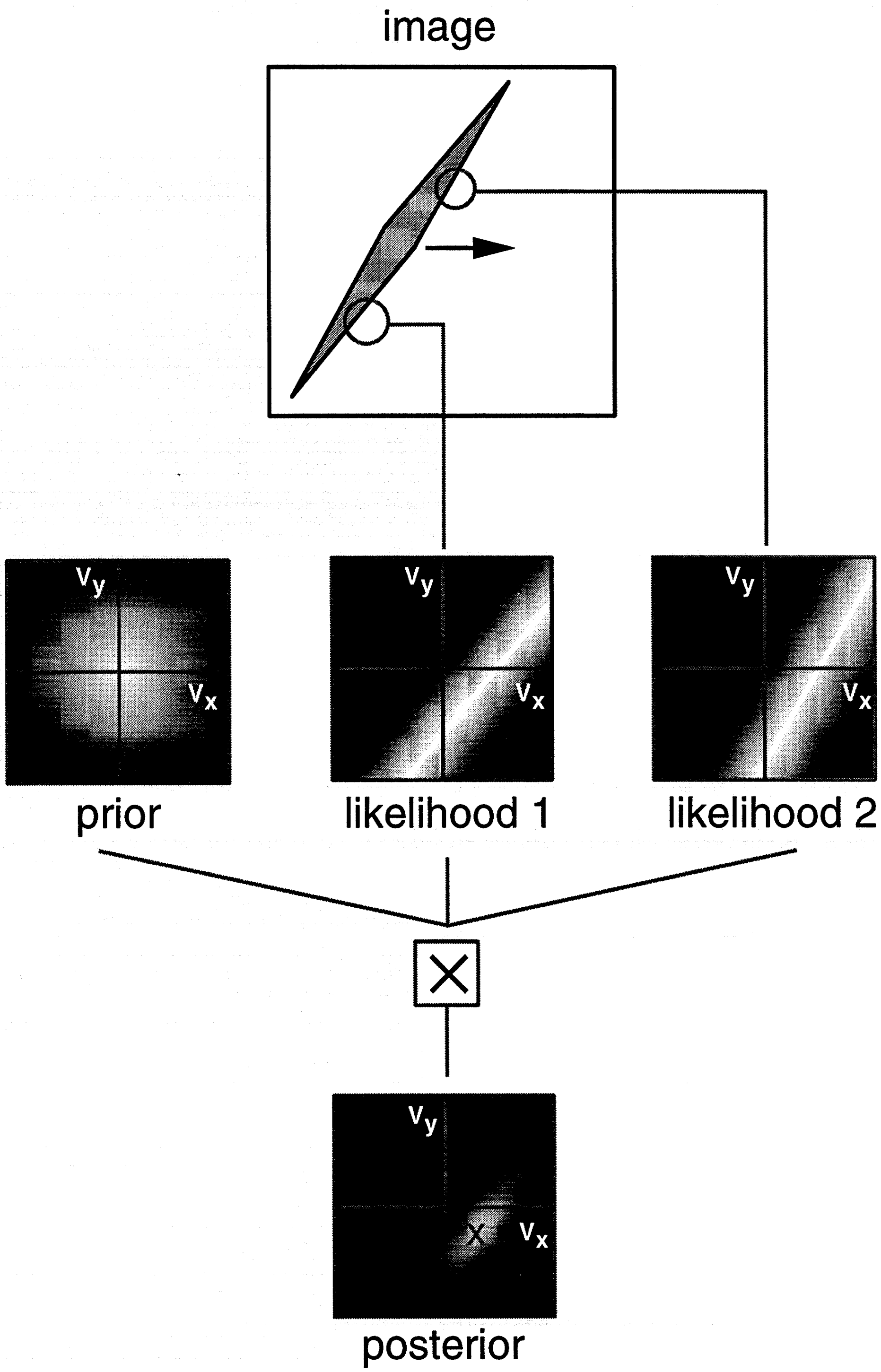


Figure 2-5: A restricted Bayesian estimator for velocity. The algorithm receives local likelihoods from various image locations and calculates the posterior probability in velocity space. This estimator is too simplistic to account for the range of phenomena we are interested in explaining but serves to give intuition about how Bayesian motion estimation works. Here the likelihoods fall off in a Gaussian manner with distance from the constraint line, and the MAP estimate is the vector average.

of phenomena we are interested in explaining. However, understanding the restricted model may help understand the more general Bayesian model.

While the Bayesian approach to perception has recently been used by a number of researchers (see e.g. (Knill and Richards, 1996)), different authors may mean different things when they refer to the visual system as Bayesian. Here we refer to two aspects of Bayesian inference - (1) that different measurements are combined while taking into account their degree of certainty and (2) that measurements are combined together with prior knowledge to arrive at an estimate.

To illustrate this definition, consider an observer who is trying to estimate the temperature outside her house. She sends out two messengers who perform measurements and report back to her. One messenger reports that the temperature is 80 degrees and attaches a high degree of certainty to his measurement, while the second messenger reports that the temperature is 60 with a low degree of certainty. The observer herself, without making any measurements, has prior knowledge that the temperature this time of the year is typically around 90 degrees. According to our definition, there are two ways in which the observer can be a non Bayesian. First, by ignoring the certainty of the two messengers and giving equal weight to the two estimates. Second, by ignoring her prior knowledge and using only the two measurements.

In order to perform Bayesian inference the observer needs to formalize her prior knowledge as a probability distribution and to ask both messengers to report probability distributions as well — the likelihoods of their evidence given a temperature. Denote by θ the unknown temperature, and E_a, E_b the evidence considered by the two messengers. The task of the Bayesian observer is to calculate the posterior probability of any temperature value given both sources of evidence:

$$P(\theta|E_a, E_b) \tag{2.1}$$

Using Bayes rule, this can be rewritten:

$$P(\theta|E_a, E_b) = kP(\theta)P(E_a, E_b|\theta) \tag{2.2}$$

where k is a normalizing constant that is independent of θ. Note that the right hand side of equation 2.2 requires knowing the joint probability of the evidence of the two messengers. Typically, neither of the two messengers would know this probability, as it requires some knowledge of the amount of information shared between them. A simplifying assumption is that the two messengers consider conditionally independent sources of evidence, in which case equation 2.2 simplifies into:

$$P(\theta|E_a, E_b) = kP(\theta)P(E_a|\theta)P(E_b|\theta) \tag{2.3}$$

Equation 2.3 expresses the *posterior* probability of the temperature as a product of the *prior* probability and the *likelihoods.* The *Maximum a posteriore (MAP)* estimate is the one that maximizes the posterior probability.

If the likelihoods and the prior probability are Gaussian distributions, the MAP estimate has a very simple form — it reduces to a weighted average of the two estimates and the prior where the weights are inversely proportional to the variances. Formally, assume $P(E_a|\theta)$ is a Gaussian with mean μ_a and variance V_a, $P(E_b|\theta)$ is a Gaussian with mean μ_b and variance V_b, and the prior $P(\theta)$ is a Gaussian with mean μ_p and variance V_p. Then θ^*, the MAP estimate is given by:

$$\theta^* = \frac{\frac{1}{V_a}\mu_a + \frac{1}{V_b}\mu_b + \frac{1}{V_p}\mu_p}{\frac{1}{V_a} + \frac{1}{V_b} + \frac{1}{V_p}} \tag{2.4}$$

Equation 2.4 illustrates the two properties of a Bayesian estimator — the two likelihoods are combined with a prior and all quantities are weighted by their uncertainty.

Motion perception can be considered in analogous terms. Suppose the observer is trying to estimate the velocity of a translating pattern. Different image locations give local readings of the motion with varying degrees of uncertainty and the observer also has some prior probability over the possible velocity. In a Bayesian estimation procedure, the observer would use the local readings in order to obtain likelihoods and then multiply these likelihoods and the prior probability to find the posterior.

This suggests the restricted Bayesian motion estimator illustrated in figures 2-3–2-5. The model receives as input likelihoods from two apertures, and multiplies them together with a prior probability to obtain a posterior probability in velocity space. Finally the peak of the posterior distribution gives the MAP estimate.

Figure 2-3 shows the MAP estimate when the two likelihoods are set to 1 for velocities on the constraint line and 0 everywhere else. The prior probability is a Gaussian favoring slow speeds (cf. (Heeger and Simoncelli, 1991)) — the probability falls off with distance from the origin. In this case, the prior probability plays no role, because when the two likelihoods are multiplied the result is zero everywhere except at the IOC solution. Thus the MAP estimate will be the IOC solution.

A second possibility is shown in figure 2-4. Here we assume that the likelihoods are zero everywhere except at velocities that are a fixed distance from the constraint line. Now when the two likelihoods are multiplied they give a diamond shaped region of velocity space in which all velocities have equal likelihood. The multiplication with the prior probability gives a "shaded diamond" posterior probability whose peak is shown with a dot. In this case the MAP estimate is the normal velocity of one of the slower grating.

A third possibility is shown in figure 2-5. Here we assume that the likelihoods are "fuzzy" constraint lines — likelihood decreases exponentially with distance from the constraint line. Now when the two likelihoods are multiplied they give rise to a "fuzzy" ellipsoid in velocity space. The IOC solution maximizes the combined likelihood but all velocities within the "fuzzy" ellipsoid have similar likelihoods. Multiplication with the prior gives a posterior probability whose peak is shown with the X symbol. In this case the MAP estimate is close to the vector average solution.

As the preceding examples show, this restricted Bayesian model may give rise to various velocity space combination rules, depending on the local likelihoods. However, as a model of human perception the restricted Bayesian model suffers from serious shortcomings:

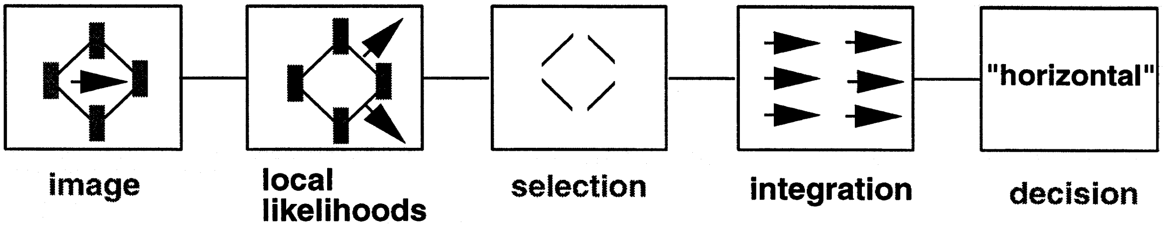


Figure 2-6: The global structure of our model. Similar to most models of motion perception, our model can be divided into two main stages - (1) a local measurement stage and (2) a global integration stage where the local measurements are combined to give an estimate of object motion. Unlike most models, the first stage extracts *probabilities* about local motion, and the second stage combines these local measurements in a Bayesian framework, taking into account a prior favoring slow and smooth velocity fields.

- The likelihood functions are based on constraint lines, i.e. on an experimenter's description of the stimulus. We need a way to calculate likelihoods directly from spatiotemporal data.
- The likelihood functions only consider "1D" locations. We need a way to define likelihoods for all image regions, including "2D" features.
- The velocity space construction of the estimator assumes rigid translation. We need a way of performing Bayesian inference for general motions, including rotations and nonrigid deformations.

In this paper we describe a more elaborate Bayesian estimator. The model works directly on the image data and combines local likelihoods with a prior probability to estimate a velocity field for a given stimulus. The prior probability favors slow and smooth velocity fields. We review a large number of previously published phenomena and find that the Bayesian estimator predicts a wide range of psychophysical results.

2.3 The model

The global structure of our model is shown in figure 2-6. As in most motion models, our model can be divided into two main stages - (1) a local measurement stage and (2) a global integration stage where the local measurements are combined to give an estimate of the motion of a surface. For present purposes we also include two stages that are not the focus of this paper - a selection stage and a decision stage.

2.3.1 Stage 1 - local likelihoods

The local measurement stage uses the output of spatiotemporal filters in order to obtain information about the motion in a small image patch. An important feature of our model is that the filter outputs are not used in order to derive a single local estimate of motion. Rather, the measurements are used to obtain a *local likelihood map* — for any particular candidate velocity we estimate the probability of the spatiotemporal data being generated by that velocity. This stage of our model is very similar to the model proposed by Heeger and Simoncelli (1991) who also suggested a physiological implementation in areas V1 and MT. Here we use a simpler, less physiological version that still captures the important notion of uncertainty in local motion measurements.

There are a number of reasons why different locations have varying degrees of ambiguity. The first reason is geometry. For a location in which the only image data is a straight edge, there are an infinite number of possible velocities that are equally consistent with the local image data (all lying on a constraint line). In a location in which the data is two-dimensional this is no longer the case, and the local data is only consistent with a single velocity.

Thus in the absence of noise, there would be only two types of measurements — "2D" locations which are unambiguous and "1D" locations which have an infinite ambiguity. However when noise is considered all locations will have some degree of ambiguity. In that case one cannot simply distinguish between velocities that "are consistent" with the local image data and those that are not. Rather the system needs to quantify the *degree* to which the data is consistent with a particular velocity.

Here we quantify the degree of consistency using the gradient constraint (Horn and Schunck, 1981; Lucas and Kanade, 1981):

$$C(v_x, v_y) = \sum_{x,y,t} w(x, y, t)(I_x v_x + I_y v_y + I_t)^2 \tag{2.5}$$

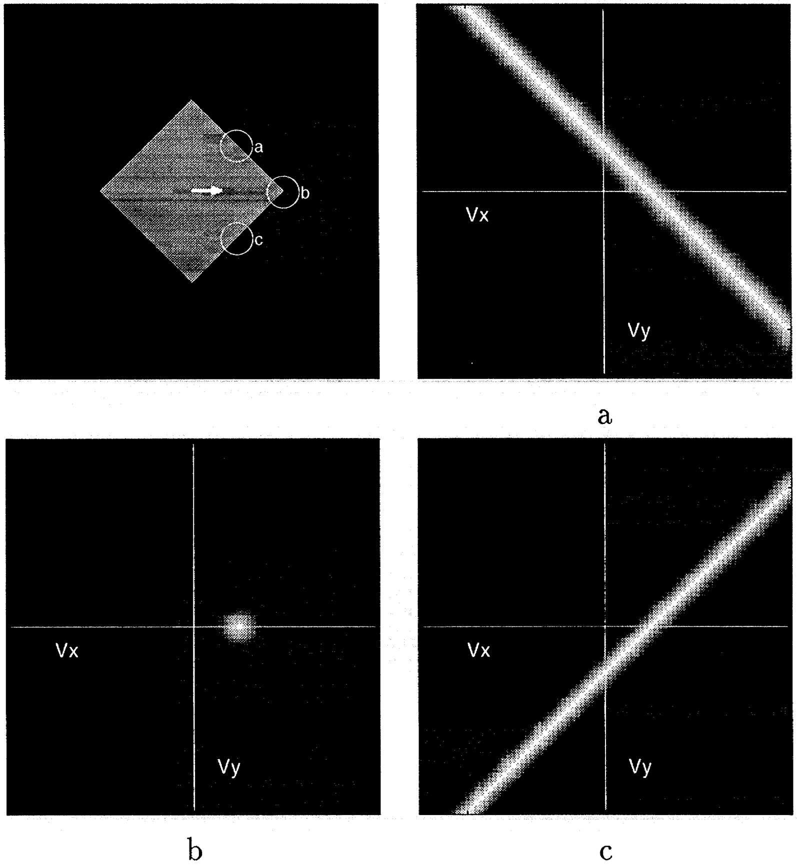
where v_x, v_y denote the horizontal and vertical components of the local velocity I_x, I_y, I_t denote the spatial and temporal derivatives of the intensity function and $w(x, y, t)$ is a spatiotemporal window centered at (x, y, t). The gradient constraint is closely related to more physiologically plausible methods for motion analysis such as autocorrelation and motion energy (Reichardt, 1961; Poggio and Reichardt, 1973; Adelson and Bergen, 1986; Simoncelli, 1993).

Assuming the intensity of a point is constant as it moves in the image the gradient constraint will be satisfied exactly for the correct velocity. If the local spatiotemporal window contains more than one orientation, the correct velocity can be determined. In the presence of noise, however, the gradient constraint only gives a relative likelihood for every velocity — the closer the constraint is to being satisfied, the more likely that velocity is. A standard derivation under the assumption of Gaussian noise in the temporal derivative (Simoncelli, 1993) gives the likelihood of a velocity at a given location:

$$L(v_x, v_y) = P(I_x, I_y, I_t | v_x, v_y) = \alpha e^{-C(v_x, v_y)/2\sigma^2} \tag{2.6}$$

where α is a normalizing constant and σ^2 is the expected variance of the noise in the temporal derivative. This parameter is required in order to convert from the consistency measure to likelihoods. If there is no noise at all in the sequence, then any small deviation from the gradient constraint for a particular velocity means that velocity is extremely unlikely. For larger amounts of noise, the system can tolerate larger deviations from the gradient constraint.

To gain intuition about the local likelihood, we display it as a gray level image for several simple stimuli (figures 2-7–2-10). In these plots the brightness at a pixel is proportional to the likelihood of a particular local velocity hypothesis - bright pixels correspond to high likelihoods while dark pixels correspond to low likelihoods.



a

b

c

Figure 2-7: A single frame from a sequence in which a diamond translates horizontally. **a-c.** Local likelihoods at three locations. At an edge the local likelihood is a “fuzzy” constraint line, while at corners the local likelihood is peaked around the veridical velocity. In this paper we use the gradient constraint to calculate these local likelihoods but very similar likelihoods were calculated using motion energy in (Simoncelli, 1993)

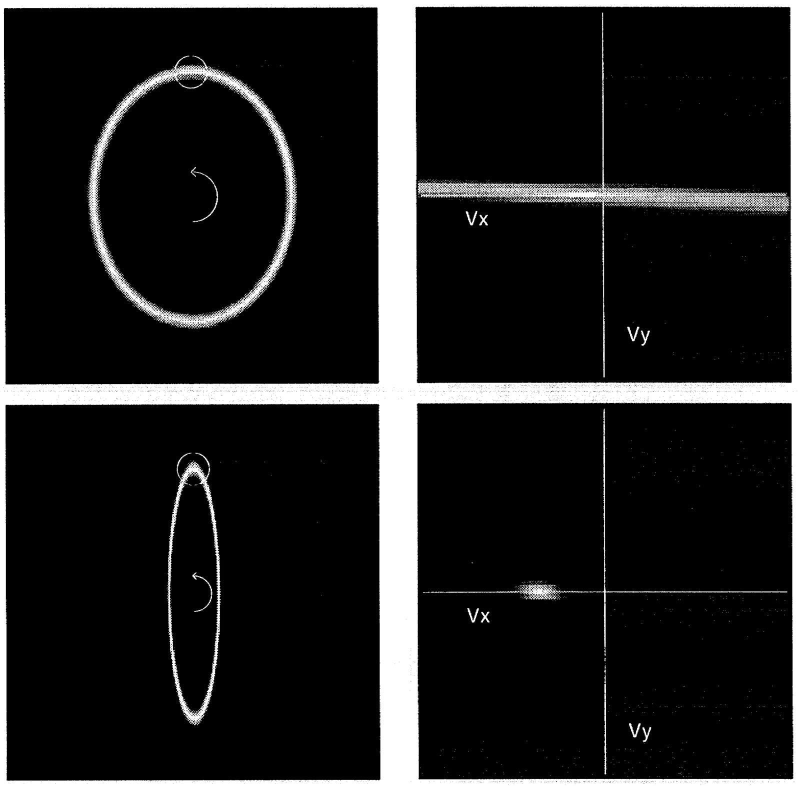


Figure 2-8: When a curved object rotates, the local information has varying degrees of ambiguity regarding the true motion, depending on the shape. In a "fat" ellipse, the local likelihood at the top of the ellipse is highly ambiguous, almost as in a straight line. In a "narrow" ellipse, however, the local likelihood at the top of the ellipse is relatively unambiguous.

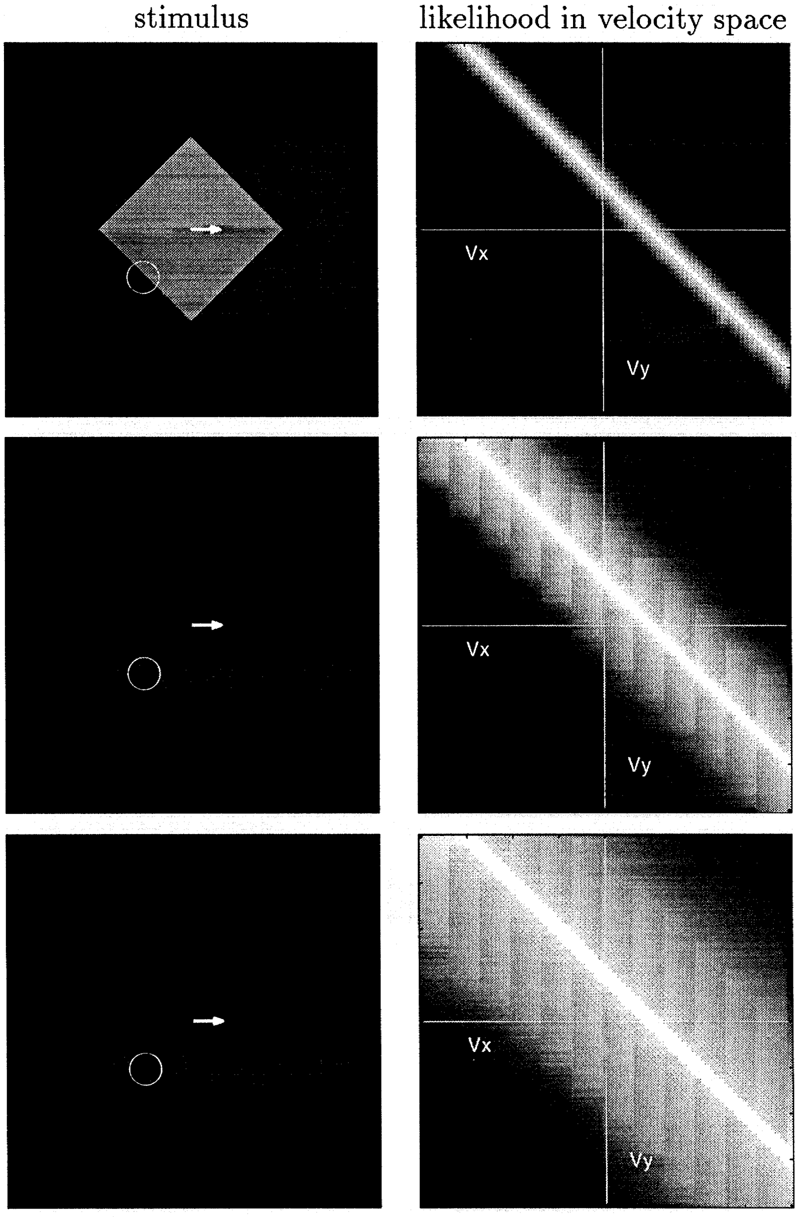


Figure 2-9: The effect of contrast on the local likelihood. As contrast decreases the likelihood becomes more fuzzy. (cf. (Simoncelli, 1993))

.

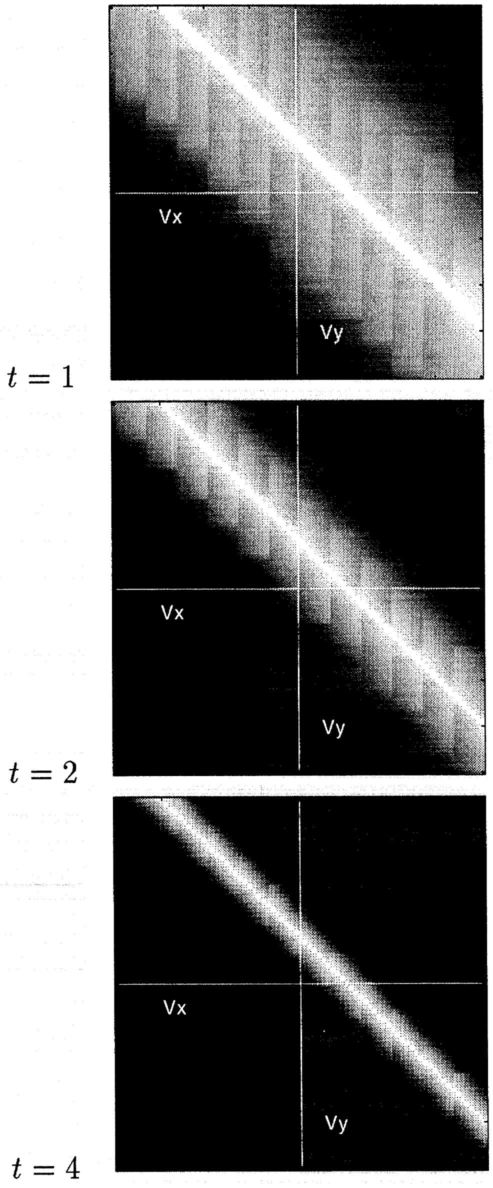


Figure 2-10: The effect of duration on the local likelihood. As duration increases the likelihood becomes more peaked.

Figure 2-7a illustrates the likelihood function at three different receptive fields on a diamond translating horizontally. Note that for locations which have straight lines, the likelihood function is similar to a "fuzzy" constraint line - all velocities on the constraint line have highest likelihood and it decreases with distance from the line. The "fuzziness" of the constraint line is governed by the parameter σ - if we assume no noise in the sequence, $\sigma = 0$, then all velocities off the constraint line have zero, but if we assume noise the falloff is more gradual and points off the constraint line may have nonzero probability. Note also that at corners where the local information is less ambiguous, the likelihood no longer has the elongated shape of a constraint line but rather is centered around the veridical velocity. Our model does not categorize locations into "corners" versus "lines" – all image locations have varying degrees of ambiguity. Figure 2-8 illustrates the likelihoods at the top of a rotating ellipse. In a "fat" ellipse, the local likelihood at the bottom of the ellipse is highly ambiguous, almost as in a straight line. In a "narrow" ellipse, however, the local likelihood at the bottom of the ellipse is highly unambiguous.

In addition to the local geometry, the uncertainty associated with a location varies with contrast and duration. Although the true velocity will always exactly satisfy the gradient constraint, at low contrasts it will be difficult to distinguish the true velocity from other candidate velocities. The degree of consistency of all velocities will be nearly identical. Indeed in the limiting case of zero contrast, there is no information at all about the local velocity and there is infinite uncertainty. Figure 2-9 shows the change in the likelihood function *for a fixed* σ as the contrast is varied. At high contrasts the likelihood function is a relatively sharp constraint line, but at lower contrasts it becomes more and more fuzzy — the less contrast the higher the uncertainty. This dependence of uncertainty on contrast is not restricted to the particular choice of consistency measure. Similar plots were obtained using motion energy in (Simoncelli, 1993).

Similarly, the shorter the duration of the stimulus the higher the uncertainty. Since the degree of consistency is summed over space and time, it is easier to distinguish the correct velocity from other candidates as the duration of the stimulus increases.

Figure 2-10 illustrates this dependence - as duration increases there is more information in the spatiotemporal receptive field and hence less uncertainty. The likelihood function becomes less fuzzy as duration increases. The quantitative dependence will of course vary with the size and the shape of the window function $w(x, y, t)$, but the

We emphasize again that in the first stage no decision is made about the local velocity. Rather in each local region, a probability distribution summarizes the range of possible velocities consistent with the local data, and the relative likelihood of each of these velocities. The combination of these local likelihoods are left to subsequent processing.

2.3.2 Stage 2 - Bayesian combination of local signals

Given the local measurements obtained across the image, the second stage calculates the MAP estimate for the motion of a single surface. In the restricted Bayesian model discussed in the introduction, this calculation could be easily performed in velocity space — it required multiplying the likelihoods and the prior to obtain the posterior.

When we consider general motions of a surface, however, the velocity space representation is not sufficient. Any $2D$ translation of a surface can be represented by a single point in velocity space with coordinates (v_x, v_y). However, there is no way to represent a rotation of a surface in a single velocity space plot, we need a larger, higher dimensional space. Figure 2-11 shows a simple generalization in which motion is represented by three numbers — two translation numbers and a rotation angle. This space is rich enough to capture rotations, but again is not rich enough to capture the range of surface motions — there is no way to capture expansion, shearing or nonrigid deformation. We need a yet higher dimensional space.

We use a 50 dimensional space to represent the motion of a surface. The mapping from parameter space to the velocity field is given by:

$$v_x(x, y) = \sum_{i=1}^{25} \theta_i G(x - x_i, y - y_i) \qquad (2.7)$$

$$v_y(x, y) = \sum_{i=26}^{50} \theta_i G(x - x_i, y - y_i) \qquad (2.8)$$

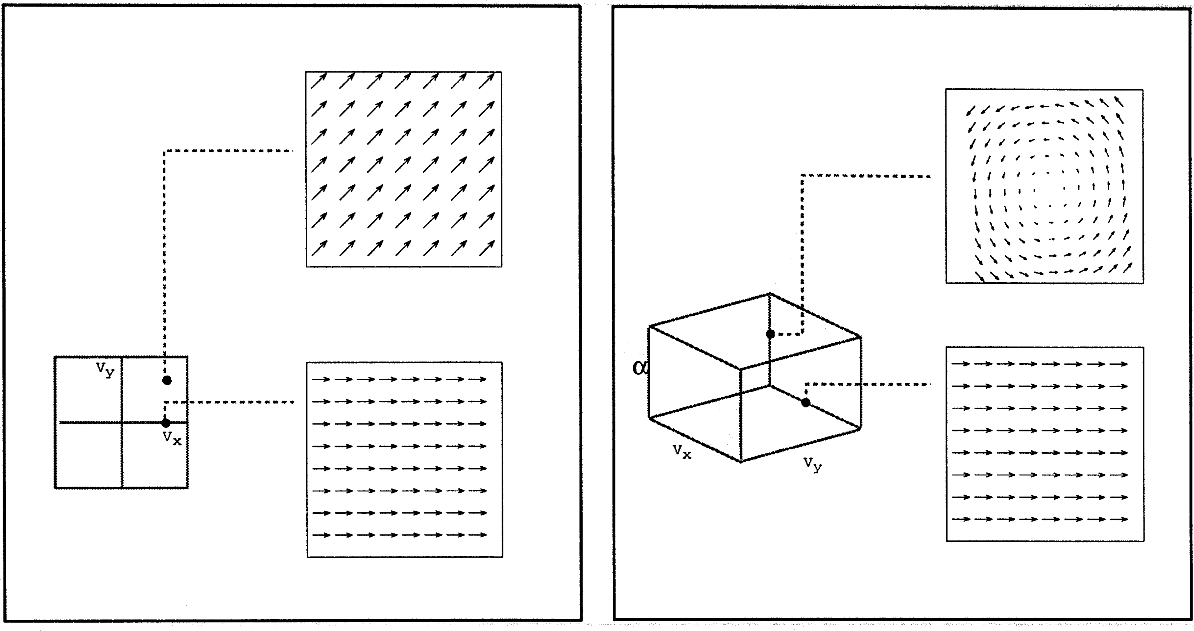


Figure 2-11: Parametric description of velocity fields. The two dimensional velocity space representation can only represent translational velocity fields. A three dimensional space can represent translational and rotational velocity fields. In this paper we use a 50 dimensional space to represent a rich family of motions including rigid and nonrigid velocity fields.

where $\{x_i, y_i\}$ are 25 locations in the image equally spaced on a $5x5$ grid and $G(x,y)$ is a two dimensional Gaussian function in image space, with spatial extent defined by σ_x:

$$G(x,y) = e^{-\frac{x^2+y^2}{2\sigma_x^2}} \tag{2.9}$$

There is nothing special about this particular representation — it is merely one choice that allows us to represent a large family of motions with a relatively small number of dimensions. We have also obtained similar results on a subset of the phenomena discussed here with other, less rich, representations.

As in the restricted Bayesian model, we need to define a prior probability over the velocity fields. This is a crucial part of specifying a Bayesian model - after all, one can make a Bayesian model do anything by designing a sufficiently complex prior. Here we choose a simple prior and show how it can account for a wide range of perceptual phenomena.

Our prior incorporates two notions: slowness and smoothness. Suggestions that

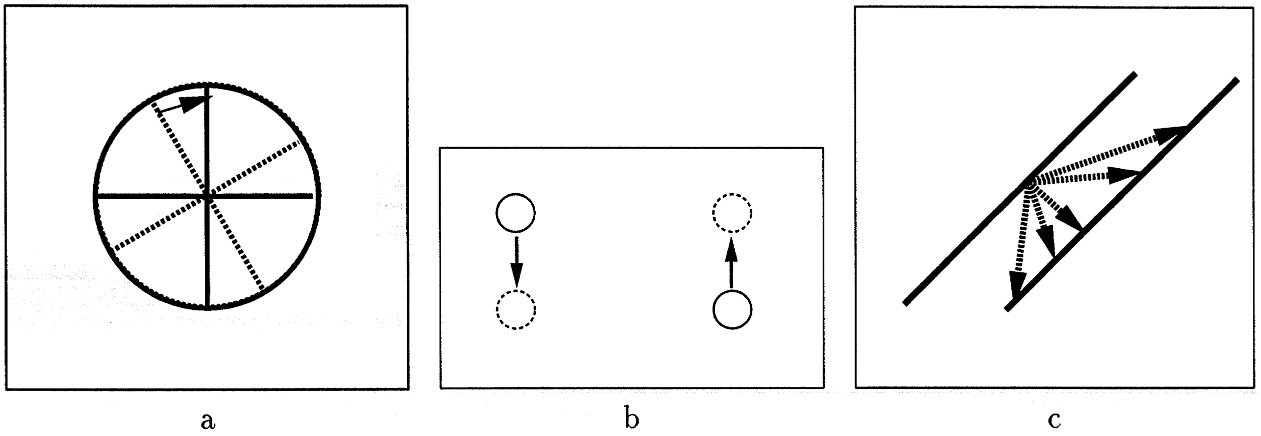


Figure 2-12: Examples of the preference for slow motions. **a.** A temporally sampled wagonwheel appears to rotate in the shortest direction. **b.** In the "quartet" stimulus, horizontal or vertical motion is perceived depending on which is shortest. **c.** A line whose endpoints are invisible is perceived as moving in the normal, or shortest, velocity.

humans tend to choose the "shortest path" or "slowest" motion consistent with the data date back to the beginning of the century (see (Ullman, 1979) and references within). Figure 2-12a shows two frames of an apparent motion stimulus. Both horizontal and vertical motions are consistent with the information but subjects invariably choose the shortest path motion. Similarly in figure 2-12b, the wagon wheel may be moving clockwise or counterclockwise but subjects tend to choose the "shortest path" or slower motion. Figure 2-12c shows an example from continuous motion. The motion of a line whose endpoints are occluded is consistent with an infinite family of velocities, yet subjects tend to prefer the normal velocity, which is the slowest velocity consistent with the data (Wallach, 1935).

However, if taken by itself, the bias towards slow speeds would lead to highly nonrigid motion percepts in curved objects. For any image sequence, the slowest velocity field consistent with the image data is one in which each point along a contour moves in the direction of its normal, and hence for objects this would predict nonrigid percepts. A simple example is shown in figure 2-13 (after Hildreth, 1983). A circle translates horizontally. The slowest velocity field is shown in figure 2-13b and is highly nonrigid. Hildreth and others (Hildreth, 1983; Horn and Schunck, 1981; Poggio et al., 1985) have therefore suggested the need for a bias towards "smooth" velocity

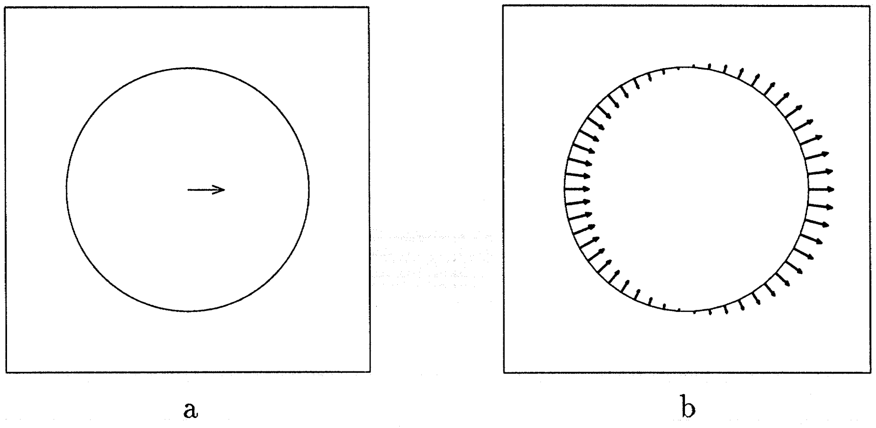


Figure 2-13: Example of the preference for smooth motions. (after (Hildreth, 1983)) **a.** A horizontally translating circle. **b.** The slowest velocity field consistent with the stimulus. Based only on the preference towards slower speeds, this stimulus would appear to deform nonrigidly.

fields, i.e. ones in which adjacent locations in the image have similar velocities.

To combine the preferences towards (1) slow and (2) smooth motions, we define a prior probability on velocity fields that penalizes for (1) the speed of the velocities and (2) the magnitude of the derivatives of the velocities. Both of these "costs" are summed over the extent of the image. The probability of the velocity field is inversely proportional to the sum of these costs. Thus the most probable velocity field is one in which the surface is static – both the speed and the derivatives of the velocity field are everywhere zero. Velocity fields corresponding to rigid translation in the image plane will also have high probability — since the velocity is constant as a function of space, the derivatives will be everywhere zero. In general, for any candidate velocity field that can be parameterized by $\vec{\theta}$ we can calculate the prior probability.

Formally, we define the following prior on a velocity field, $V(x, y)$:

$$P(V) = \alpha e^{-J(V)} \tag{2.10}$$

with:

$$J(V) = \sum_{xy} \|Dv(x, y)\|^2 \tag{2.11}$$

here Dv is a differential operator, i.e. it measures the derivatives of the velocity field.

We follow Grzywacz and Yuille (1991) in using a differential operator that penalizes velocity fields with strong derivatives:

$$Dv = \sum_{n=0}^{\infty} a_n \frac{\partial^n}{\partial x} v \tag{2.12}$$

Note that the sum starts from $n = 0$ thus Dv also includes a penalty for the "zero order" derivative - i.e. it penalizes fast flow fields. For mathematical convenience, Grzywacz and Yuille chose $a_n = \lambda^{2n}/(n!2^n)$ where λ is a free parameter. They noted that similar results are obtained when a_n is set to zero for $n > 2$. We have also found this to be true in our simulations. Thus the main significance of the parameter λ is that it controls the ratio between the penalty for fast velocities ($a_0 = 1$) and the penalty for nonsmooth velocities ($a_1 = \lambda^2/2$). We used a constant value of λ throughout (see appendix).

Unlike the restricted Bayesian model discussed in the introduction, the calculation of the posterior probability cannot be performed graphically. The prior probability of $\vec{\theta}$ for example is a probability distribution over a 50 dimensional space. However, as we show in the appendix it is possible to solve analytically for the most probable $\vec{\theta}$. This gives the velocity field predicted by the model for a given image sequence.

2.3.3 Selection and Decision

As mentioned in the introduction, in scenes containing multiple objects, the selection of which signals to integrate is a crucial step in motion analysis (cf. (Nowlan and Sejnowski, 1995)). This is not the focus of our paper, but in order to apply our model directly to raw images we needed some rudimentary selection process. We make the simplifying assumption that the image contains a single moving object and (optionally) static occluders. Thus our selection process is based on subtracting subsequent frames and thresholding the subtraction to find regions that are not static. All measurements from these regions are combined. The selection stage also discards all measurements from receptive fields lying exactly on the border of the image, to avoid edge artifacts.

The decision stage is needed in order to relate our model to psychophysical experiments. The motion integration stage calculates a velocity field, but in many experiments the task calls for making a discrete decision based on the perceived velocity field (e.g. "up" versus "down"). In order to model these experiments, the decision stage makes a judgment based on the estimated velocity field. For example, if the experiment calls for a direction of motion judgment, the decision stage fits a single global translation to the velocity field and output the direction of that translation.

2.3.4 Model Summary

The model starts by obtaining local velocity likelihoods at every image location. These likelihoods are then combined in the second stage to calculate the most probable velocity field, based on a Bayesian prior favoring slow and smooth motions. All results described in the next section were obtained using the Gaussian parameterization (equation 2.7), with a fixed λ. Stimuli used as input were gray level image sequences (5 frames $128x128$ pixel size) and the spatiotemporal window used to calculate the likelihoods was of size $5x5x5$ pixels.

The only free parameter that varies between experiments is the parameter σ. It corresponds to the observer's assumption about the reliability of his or her temporal derivative estimates. Thus we would expect the numerical value of σ to vary somewhat between observers. Indeed for many of the illusions we model here, individual differences have been reported for the magnitude of the bias (e.g. (Yo and Wilson, 1992; Lorenceau et al., 1992)) although the qualitative nature of the perceptual bias is similar across subjects. Although σ is varied when modeling different experiments, it is always held constant when modeling a single experiment, thus simulating the response of a single observer to varying conditions.

2.4 Results

We start by showing the results of the model on translating stimuli. Although the Bayesian estimate is a velocity *field*, we summarize the estimate for these stimuli using a single velocity vector. This vector is calculated by taking the weighted mean value of the velocity field with weight decreasing with distance from the center of the image. Except otherwise noted the estimated velocity field is roughly constant as a function of space and is well summarized with a single vector.

2.4.1 The Barberpole illusion - Wallach 35

Phenomena: As noted by Wallach (1935), a grating viewed through a circular aperture is perceived as moving in the normal direction, but a grating viewed through a rectangular aperture is perceived as moving in the direction of the longer axis of the aperture.

Model Results: Figure 2-14b,d shows the Bayesian estimate for the two stimuli. In the circular aperture the Bayesian estimate is in the direction of the normal velocity, while in the rectangular one, the estimate is in the direction of the longer axis of the aperture.

Discussion: Recall that the Bayesian estimate combines measurements from different locations according to their uncertainty. For the rectangular aperture, the "terminator" locations corresponding to the edges of the aperture dominate the estimate and the grating is perceived to move horizontally. In the circular aperture, the terminators do not move in a coherent direction, and hence do not have a large influence on the estimate. Among all velocities consistent with the constraint line, the preference for slow speeds favors the normal velocity.

For the rectangular aperture the Bayesian estimate exhibits significant nonrigidity — at the vertical edges of the aperture the field has strong vertical components. We also note that although the present model can account for the basic barberpole effect, it does not account for various manipulations that influence the terminator classification and the magnitude of the barberpole effect. For example, Shimojo et al. (1989) have used stereoscopic depth to place the grating behind the aperture and their subjects tended to perceive the grating as moving closer to the normal direction even in a rectangular aperture. A more sophisticated selection mechanism is required to account for their effect.

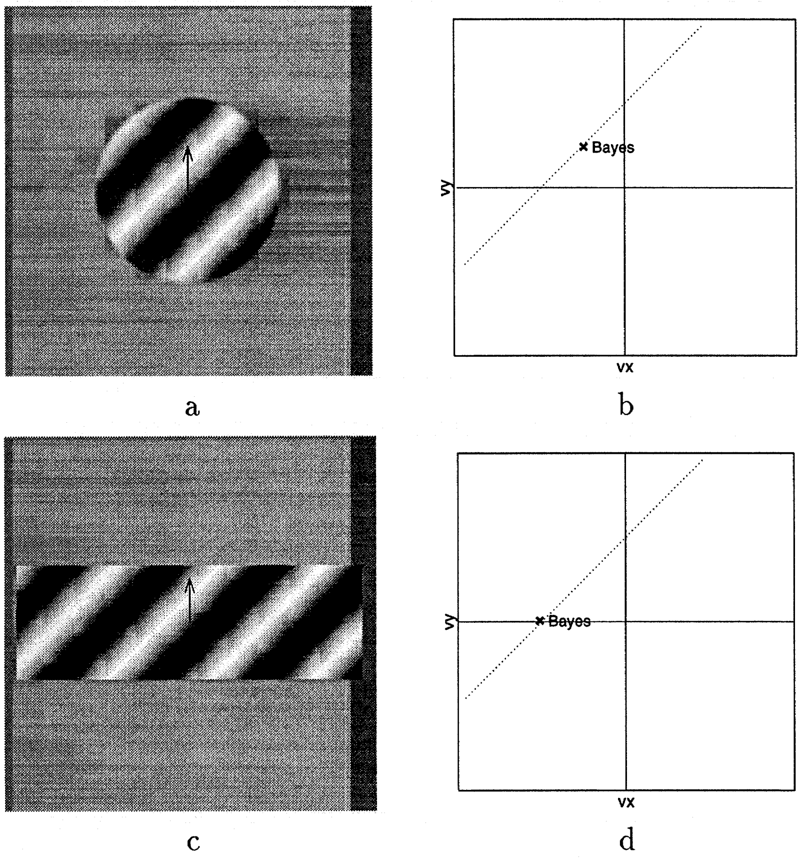


Figure 2-14: The "barberpole" illusion (Wallach 35). A grating viewed through an (invisible) circular aperture is perceived as moving in the normal direction, but a grating viewed through a rectangular aperture is perceived as moving in the direction of the long axis. **a** A grating viewed through a circular aperture. **b.** The Bayesian estimate for this sequence. Note that the Bayesian estimate is in the normal direction. **c.** A grating viewed through a rectangular aperture. **d.** The Bayesian estimate for this sequence. Note that the Bayesian estimator is now in the direction of the longer axis. Because measurements are combined according to their uncertainty, the unambiguous measurements along the aperture edge overcome the ambiguous ones obtained inside the aperture.

2.4.2 Biases towards VA in translating stimuli

Type II plaids - Yo and Wilson (1992)

Phenomena: Yo and Wilson (1992) distinguished between two types of plaid figures. In "Type I" plaids the two normal velocities lie on different sides of the veridical velocity, while in "type II" plaids both normal velocities lie on the same side and hence the vector average is quite different from the veridical velocity (see figure 2-15b,d). They found that for short presentation times, or low contrast, the perceived motion of type II is strongly biased in the direction of the vector average while the percept of type I plaids is largely veridical.

Model Results: Figure 2-15b,d shows the VA, IOC and Bayesian estimate for the two stimuli. For type I plaids the estimated direction is veridical but the speed is slightly slower than the veridical. For type II plaids the Bayesian estimator gives an estimate that is far from the veridical velocity, and that is much closer to the vector average.

Discussion: The decrease speed observed in the Bayesian estimate for type I plaids is to be expected from a prior favoring slow velocities. The bias in direction towards the VA in type II plaids is perhaps less obvious. Where does it come from?

As pointed out by Heeger and Simoncelli (1991), a Bayesian estimate with a prior favoring slow speeds will be biased towards VA in this case, since the VA solution is much slower. Consider figure 2-15b. Recall that the Bayesian estimate maximizes the product of the likelihood and the prior of the estimate. Let us compare the veridical IOC solution to the Bayesian estimate in these terms.

In terms of **likelihood** the IOC estimate is optimal. It is the only solution that lies exactly on both constraint lines. The Bayesian solution does not maximize the likelihood, since it does not lie exactly on both constraint lines. However, recall that the local likelihoods are "fuzzy" constraint lines, and hence the Bayesian solution which is close to both constraint lines still receives high likelihood. In terms of the **prior**, however, the Bayesian solution is much preferred. It is significantly (about 55%) slower than the IOC solution. Thus a system that maximizes both the prior

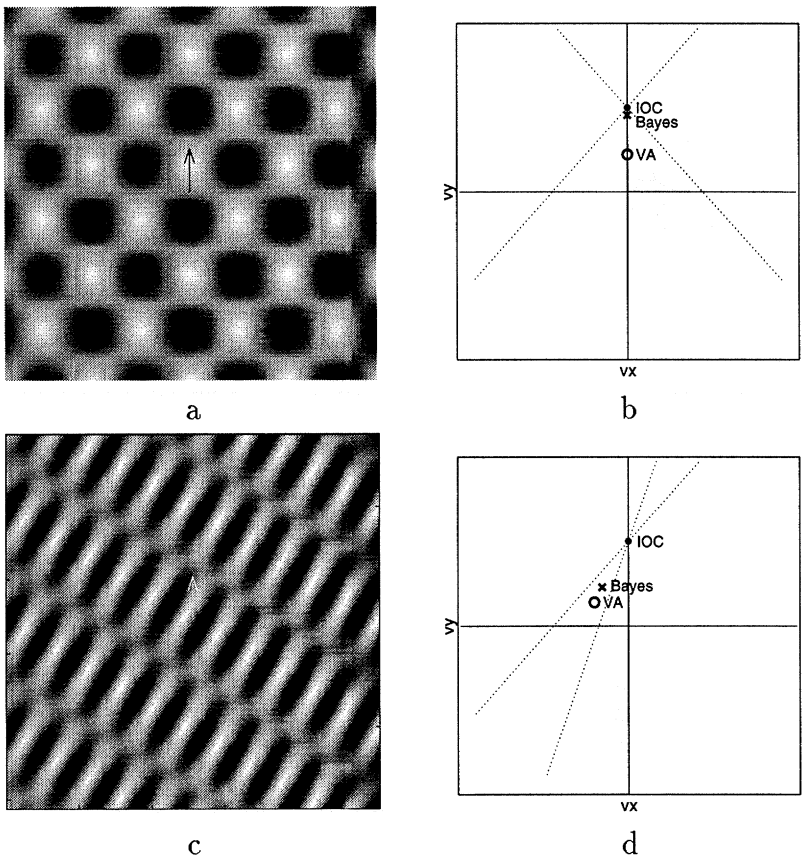


Figure 2-15: A categorization of plaid patterns introduced by Yo and Wilson (1992). “Type I” plaids have component motions on both sides of the veridical velocity, while “Type II” plaids do not. **a** a “type I” plaid moving upward is typically perceived veridically. **b.** The IOC, VA and Bayesian estimate for this sequence. Note that the Bayesian estimate is in the veridical direction. **c.** a “type II” plaid moving upward is typically perceived to move in the direction of the vector average. **d.** The IOC, VA and Bayesian estimate for this sequence. Note that the Bayesian estimator is biased towards the VA motion, as is the percept of observers. Although the IOC solution maximizes the likelihood, the VA solution has higher prior probability and only slightly lower likelihood.

and the likelihood will not choose the IOC solution, but rather one that is biased towards the vector average.

Note that this argument only holds when the likelihoods are "fuzzy" constraint lines, i.e. when the system assumes some noise in the local measurements. A system that assumed no noise would give zero probability to any velocity that did not lie exactly on both constraint lines and would always choose the IOC solution. Recall that the degree of "fuzziness" of the constraint lines varies depending on the conditions, e.g. the contrast and duration of the stimulus. Thus the Bayesian estimate may shift from the VA to the IOC solution depending on the conditions. In subsequent sections we show that to be the case.

Biased oriented lines - Mingolla et al. (1992)

Phenomena: Additional evidence for a vector average combination rule was found by Mingolla et al. (1992) using stimuli consisting of lines shown behind apertures (see figure 2-16a). Behind each aperture, a line translates horizontally, and the orientation of the line is one of two possible orientations. In the "downward biased" condition, the lines are $+15, +45$ degrees from vertical, in the "upward biased" condition, the lines are $-15, -45$ from vertical and in the "no bias" condition the lines are $+15, -15$ degree from vertical. They found that the perceived direction of motion is heavily biased by the orientation of the lines. In a two alternative forced choice experiment, the upward, downward and unbiased line patterns moved in five directions of motion. Subjects were asked to indicate whether the motion was upward or downward. Figures 2-17a shows the performance of the average subject on this task, replotted from (Mingolla et al., 1992). Note that in the biased conditions, subjects' percept is completely due to the orientation of the lines and is independent of the actual motion.

Model Results: Figure 2-16b shows the IOC, VA and Bayesian solution for the stimulus shown in figure 2-16a. The Bayesian solution is indeed biased upwards. Figure 2-17b shows the 'percent correct' of the Bayesian model in a simulated 2AFC experiment. To determine the percentage of upward responses, the decision module

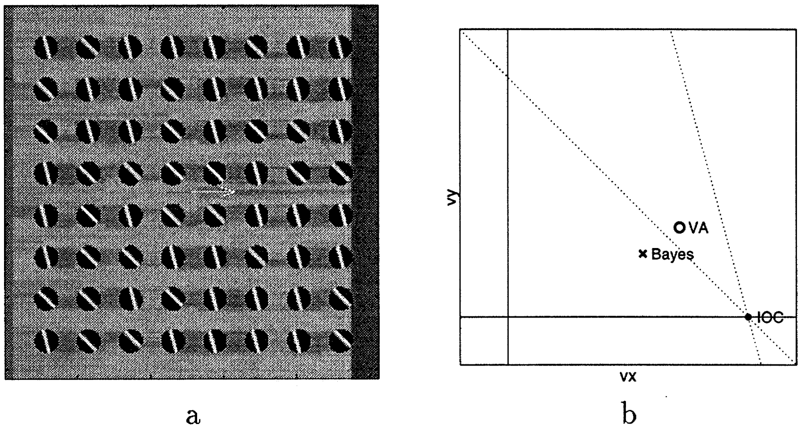


Figure 2-16: A stimulus studied by Mingolla et al. (1992) suggesting a vector average combination rule. **a.** a single frame from a sequence in which oriented lines move horizontally behind apertures. **b.** The IOC, VA and Bayesian estimate for this sequence. Note that the Bayesian estimator is biased towards the VA motion, as is the percept of observers (Mingolla et al., 1992).

used a "soft" threshold on the velocity field:

$$P = \frac{1}{1 + exp(-\alpha)} \tag{2.13}$$

where α is the model's estimated direction of motion. This corresponds to a "soft" threshold decision on the model's output. The only free parameter, σ was held constant throughout these simulations. Note that in the biased conditions, the model's percept is completely due to the orientation of the lines and is independent of the actual motion.

Discussion: As in the type II plaid, the veridical velocity is not preferred by the model, due to the prior favoring slower speeds. The veridical velocity maximizes the likelihood but not the posterior. In a second set of simulations (not shown) the terminations of the line endings were visible inside each aperture. Consistent with the results of Mingolla et al. (1992), the estimated direction was primarily a function of the true direction of the pattern and not the orientation.

A manifold of lines (Rubin and Hochstein 92)

Phenomena: Even in stimuli containing more than two orientations, the visual system

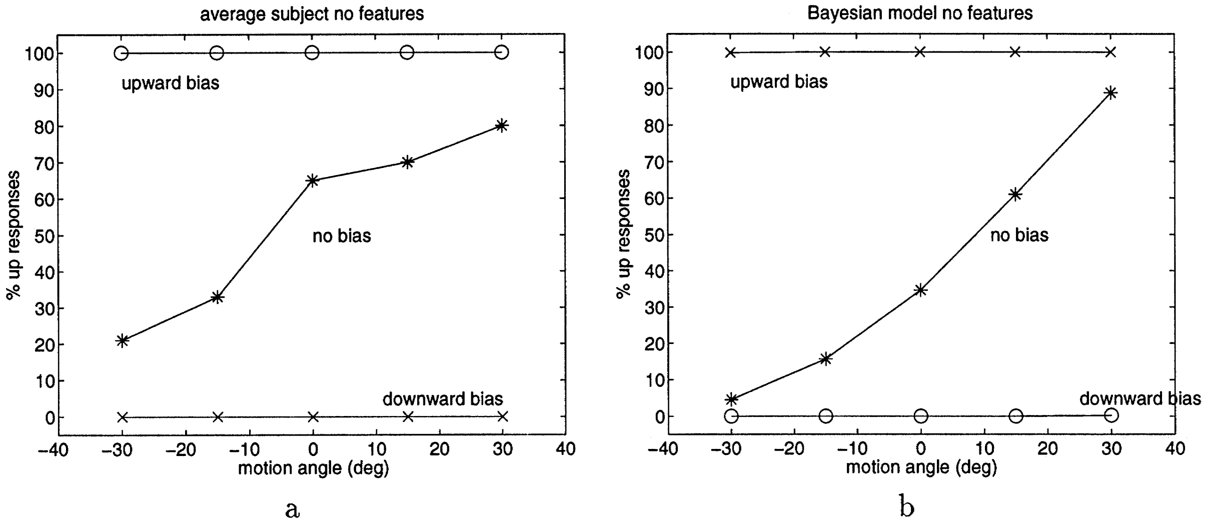


Figure 2-17: **a.** Results of experiment 1 in (Mingolla et al., 1992). Three variations on the line images shown in figure 2-16a moved in five directions of motion. Subjects were asked to indicate whether the lines moved upward or downward. Note that in the absence of features, the perceived direction was only a function of the orientation of the lines. **b.** Results of Bayesian estimator output on the same stimuli. The single free parameter σ is held constant throughout.

may be incapable of estimating the veridical velocity. Rubin and Hochstein (1993) presented subjects with a "manifold" of lines translating horizontally (see figure 2-18a). They asked subjects to adjust a pointer until it matched their perceived velocity and found that the perceived motion was diagonal, in the direction of the vector average. The authors also noted that when a small number of horizontally translating dots were added to the display (figure 2-18c), the veridical motion was perceived.

Model Results: Figure 2-18b shows the IOC, VA and Bayesian solution for the manifold stimulus. The Bayesian estimate is biased in the direction of the VA. Figure 2-18d shows the estimate when a small number of dots are added. The estimate is now veridical.

Discussion: The bias in the absence of features is explained in the previous displays — the veridical velocity maximizes the likelihood but not the posterior. The shift in percept based on a small number of terminator signals falls naturally out of the Bayesian framework. Since individual measurements are combined according to their uncertainty, the small number of measurements from the dots overcome the

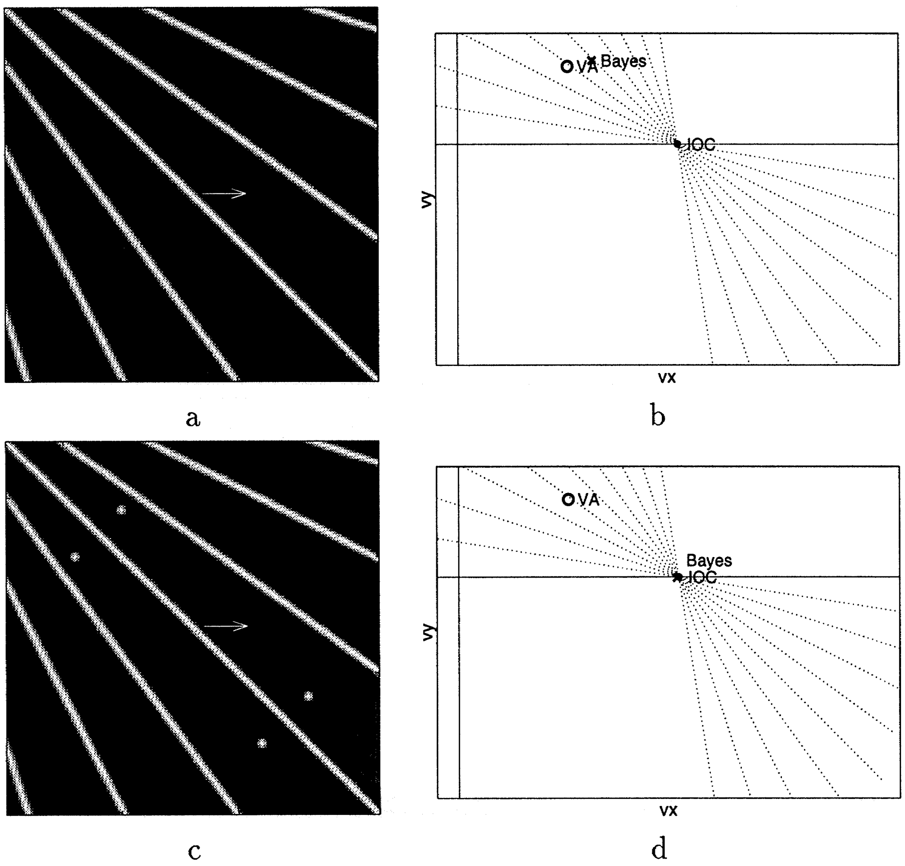


Figure 2-18: **a.** A single frame from a stimulus introduced by Rubin and Hochstein (1993). A collection of oriented lines translate horizontally. **b.** The VA, IOC and Bayesian estimate. The Bayesian estimate is biased in the vector average direction, consistent with the percept of human subjects. **c.** When a small number of dots are added to the display the pattern appears to translate horizontally (Rubin and Hochstein 92). **d.** The Bayesian estimate shifts to veridical under these circumstances. Since individual measurements are combined according to their uncertainty, the small number of measurements from the dots overcome the measurements from the lines.

measurements from the lines.

In Rubin and Hochstein's original displays the lines were viewed through an aperture, unlike the displays used here where the lines fill the image. An interesting facet of Rubin and Hochstein's results which is not captured in our model is that the accidental terminator signals created by the aperture also had a significant effect on the perceived motion. Similar to the results with the barber pole illusion, they found that manipulating the perceived depth of the aperture changed the influence of the terminators. A more sophisticated selection mechanism is needed to account for these results.

Intermediate solutions - Bowns (1996)

Phenomena: The Bayesian estimator generally gives a velocity estimate somewhere between "pure" vector average and "pure" IOC. Evidence against either pure mechanism was recently reported by Bowns (1996). In her experiment, a set of type II plaids consisting of orientations 202 and 225 were used as stimuli. Although the two orientations were held constant, the relative speeds of the two components were varied. The result was a set of plaids where the vector average was always right of the vertical while the IOC solution was always left of vertical. Figure 2-19 shows examples of the two extreme plaids used in her study, along with their velocity space construction.

Subjects were asked to determine whether or not the motion was left or right of vertical. It was found that when the speeds of the two components were similar, subjects answered right of vertical (consistent with the VA solution) while when the speeds were dissimilar subjects answered left of vertical (consistent with the VA solution). The circles in figure 2-20 show the percentage of rightward results for a subject in her experiment.

Model Results Figure 2-19c and d show the Bayesian estimate for the two extreme cases. Note that they switch from left of vertical to right of vertical as the relative speeds change. In figure 2-20 the solid line gives the expected percent rightward responses for the Bayesian estimator. Note that it gives a gradual shift from left to

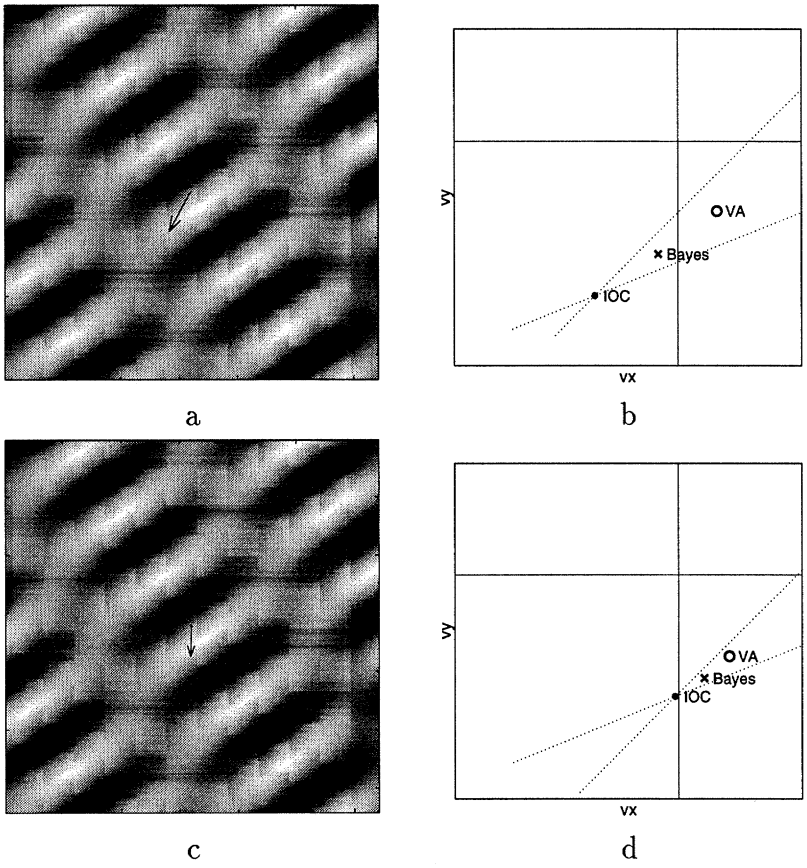


Figure 2-19: Experimental stimuli used by Bowns (1996) that provide evidence against a pure vector average or IOC mechanism. **a.** A type II plaid with orientations 202, 225 degrees and relative speeds 1, 0.45. **b.** The VA, IOC and Bayesian estimates. The IOC solution is leftward of the vertical while the VA solution is rightward. The Bayesian estimate is leftward, consistent with the results of Bowns (1996). **c.** A type II plaid with orientations 202, 225 degrees and relative speeds 1, 0.75. **d.** The VA, IOC and Bayesian estimates. The IOC solution is leftward of the vertical while the VA solution is rightward. The Bayesian estimate is rightward, consistent with the results of Bowns (1996).

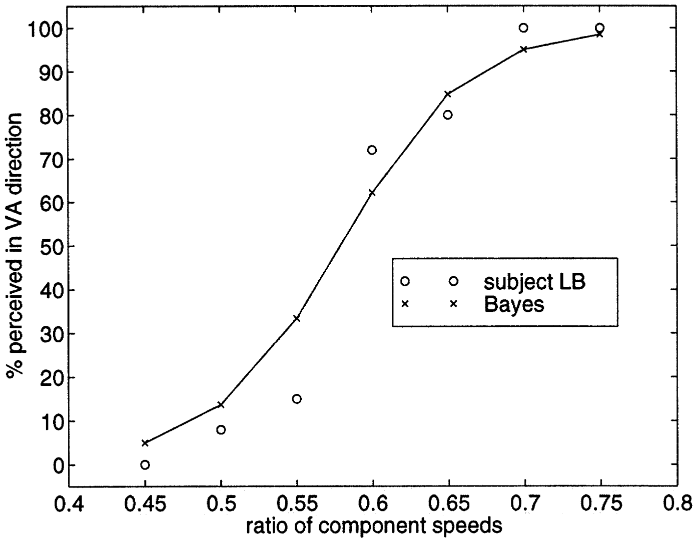


Figure 2-20: The results of an experiment conducted by Bowns (1996). Subjects indicated if the motion of a plaid was left of vertical (consistent with VA) or rightwards of vertical (consistent with IOC). The relative speeds of the two components were varied. The circles show the results of subject LB, while the crosses show the output of the Bayesian model (σ constant throughout). The experimental results are inconsistent with pure VA or pure IOC but are consistent with a Bayesian estimator.

right as the relative speeds are varied. The parameter σ is held constant throughout.

Discussion: Here again, the prior favoring slower speeds causes the Bayesian estimator to move away from the veridical IOC solution. However, the Bayesian estimator is neither a "pure" IOC solution nor a "pure" VA solution. Rather it may give any perceived velocity that varies smoothly with stimulus parameters.

The fact that a Bayesian estimator is biased towards the vector average solution suggests that the VA bias is not a result of the inability of the visual system to correctly solve for the IOC solution, but rather may be a result of a combination rule that takes into account noise and prior probabilities to arrive at an estimate.

2.4.3 Dependence of VA bias on stimulus orientation

Effect of component orientation - Burke and Wenderoth (1992)

Phenomena: Even in type II plaids, the perceived direction may be more consistent with IOC than VA (Bowns, 1996; Burke and Wenderoth, 1993). Consider, for exam-

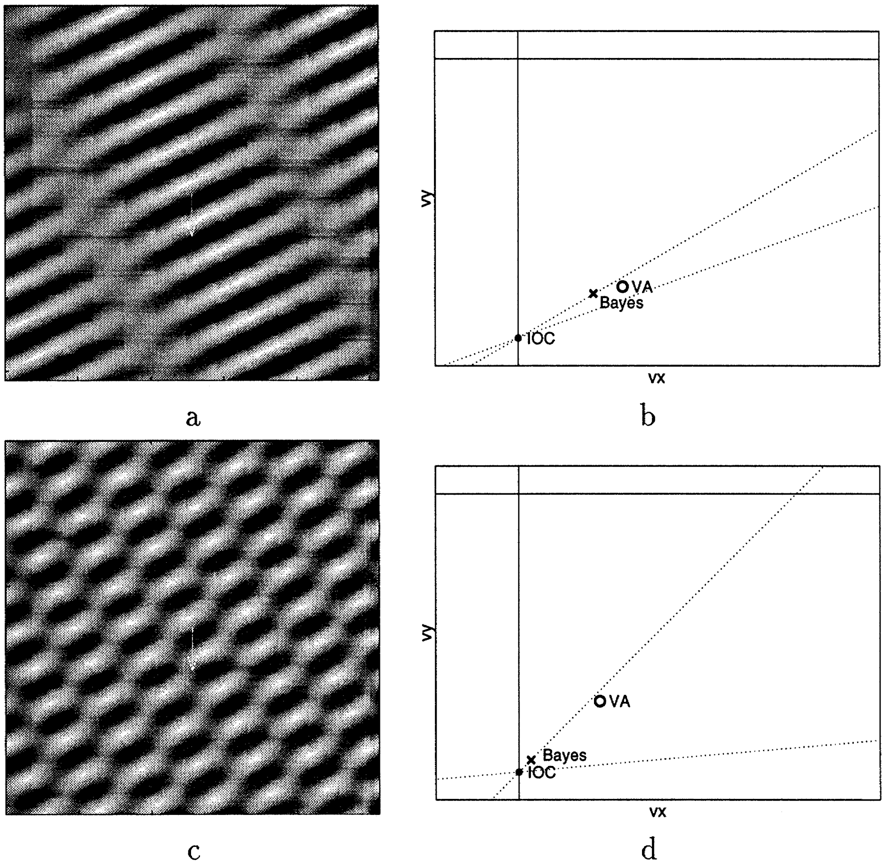


Figure 2-21: Stimuli used by Burke and Wenderoth (1993) to show that the percept of some type II plaids is more consistent with IOC than with VA. **a.** A type II plaid with orientations 20 and 30 degrees is misperceived by about 15 degrees.(Burke and Wenderoth, 1993) **b.** The VA, IOC and Bayesian estimates. The Bayesian estimate is biased in a similar manner to the human observers. **c.** A type II plaid with orientations 5 and 45 degrees is is perceived nearly veridically. (Burke and Wenderoth, 1993) **d.** The VA, IOC and Bayesian estimate. The Bayesian estimate is nearly veridical. The parameter σ is held constant throughout.

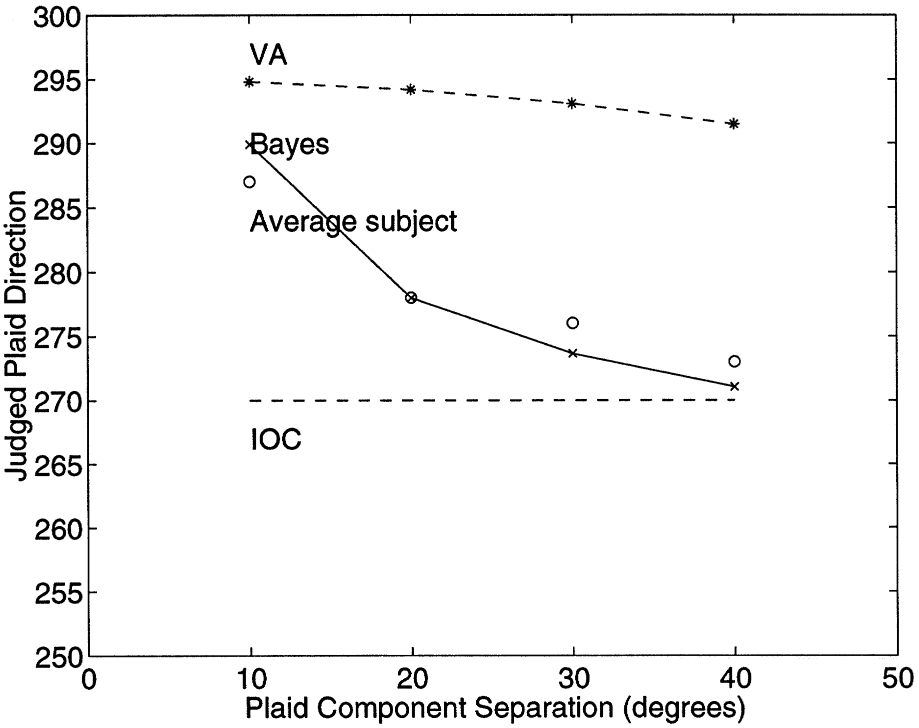


Figure 2-22: Results of an experiment conducted by Burke and Wenderoth (1993) to systematically investigate the effect of plaid component orientation on perceived direction. All they plaids are "type II" and yet when the relative angle between the components of the plaid is increased varied, the perceived direction shows a gradual shift from the VA to the IOC solution (open circles replotted from (Burke and Wenderoth, 1993)). The Bayesian estimator, with a fixed σ shows the same behavior

ple, the type II plaids shown in figure 2-21. Burke and Wenderoth (1993) found that for the plaid in figure 2-21a (orientations 200, 210) the perceived direction is biased by about 15 degrees, while for the plaid in figure 2-21c (orientations 185, 225) the perceived direction is nearly veridical with a bias of under 2 degrees. Thus if one assumes independent IOC and VA mechanisms, one would need to assume that the visual system uses the IOC mechanism for the plaid in figure 2-21c but switches to the VA mechanism for the plaid in figure 2-21a. Burke and Wenderoth systematically varied the angle between the two plaid components and asked subjects to report their perceived directions. The results are shown in the open circles in figure 2-22. The perceived direction is inconsistent with a pure VA mechanism or a pure IOC mechanism. Rather it shows a gradual shift from the VA to the IOC solution as the angle between the components increases.

Model Results: Figure 2-22 shows the predicted IOC, VA and Bayesian estimates as the angles are varied. The parameter σ is held fixed. Note that a single model generates the range of percepts, consistent with human observers.

Discussion: To get an intuitive understanding of why the same Bayesian estimator gives IOC or VA type solutions depending on the orientation of the components, compare figure 2-21b to figure 2-21d. Note that in figure 2-21b the two constraint lines are nearly parallel. Hence, a solution lying halfway between the two constraint lines (such as the VA solution) receives high likelihood for fuzzy constraint lines. However, in figure 2-21d, where the components have a 40 degrees difference in orientation, the two constraint lines are also separated by 40 degrees. Thus for a solution to have high likelihood, it is forced to lie close to the intersection of the two lines, or the IOC solution. Thus a single, Bayesian mechanism predicts a gradual shift from VA to IOC as the orientation of the components is varied.

An alternative explanation of the shift in perceived direction was suggested by Bowns (1996) who pointed out that there exist features in the "blob" regions of these plaids that move in different directions as the orientations of the gratings are varied. Our results do not of course rule out this explanation, but they show that hypothesizing a specialized "blob" mechanism is not necessary.

Orientation effects in occluded stimuli

Phenomena: We have performed experiments with the stimulus shown in figure 2-23a. A rhombus whose four corners are occluded is moving horizontally. Note that there are no features on this stimulus which move horizontally - the two normal velocities are diagonal and the terminator motion is downward. This stimulus is similar to a type II plaid in the sense that the two normal velocities lie on the same side of the veridical velocity. However it requires integration across space rather than across multiple orientations at a single point. We wanted to see whether the biases in perceived velocity would behave the same way as in plaids.

We presented subjects with these stimuli while varying the angle of one of the sides and asked them to indicate the perceived direction. Results of a typical subject are shown in figure 2-23. Consistent with the result on plaids (Burke and Wenderoth, 1993) subjects percept shift gradually from a bias in the VA direction to the veridical direction as the angular difference increases.

Model Results: Figure 2-23 shows the result of the Bayesian estimator with fixed σ. Similar to the results with plaids, the Bayesian estimate shifts gradually from a bias in the VA direction to the veridical direction as the angular difference increases.

Discussion: It seems difficult to reconcile these results with a "multiple mechanism" model in which the visual system uses a VA mechanism or an IOC mechanism depending on the conditions. First, one would have to assume that the visual system uses a different mechanism for nearly identical stimuli, when the relative orientations is changed. Second, the perceived direction changes continuously and includes intermediate values that are inconsistent with either VA or IOC.

Likewise, these results are difficult to reconcile with a "feature tracking" explanation of the sort proposed by Bownes (1996) or by Yo and Wilson (1992) . No matter what the orientation of the rhombus sides are, there are never any trackable features moving in the veridical direction. Yet subjects perceive motion in the veridical direction when the angle between the two components is large.

In contrast, as we have shown, these results are consistent with a Bayesian esti-

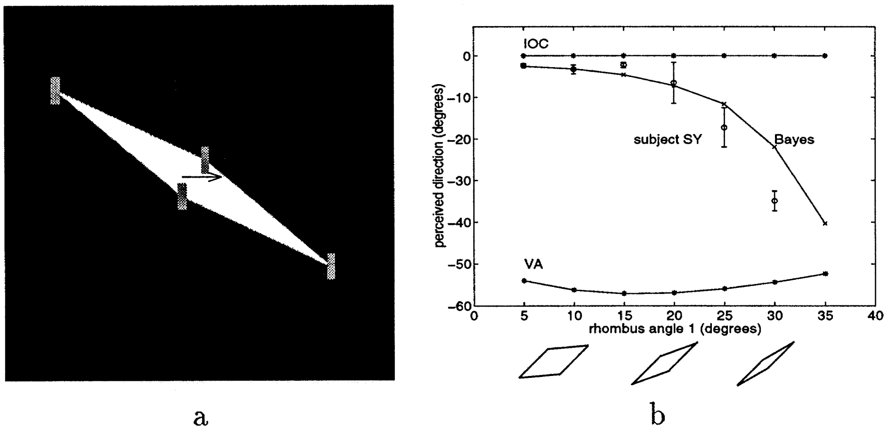


Figure 2-23: The stimulus used in an experiment to measure influence of relative orientation on perceived direction. A rhombus whose four corners are occluded was translating horizontally. The angle between the two orientations was varied. **a.** A single frame from the sequence. **b.** The predictions of VA, IOC and the Bayesian estimator for the direction of motion of the rhombus. One of the orientations is fixed at 40 degrees, and the second orientation is varied. The VA solution is always far from horizontal (by at least 50 degrees), the IOC prediction is always horizontal and the Bayesian estimator predicts a gradual shift from horizontal to diagonal as the angle between the two components is decreased. The results of a single subject are shown in circles.

mation strategy where motion signals are fused in accordance with their uncertainty and combined with a prior favoring slow and smooth velocities. Again, this does not rule out the "multiple mechanism" explanation, but shows that it is not necessary. A single Bayesian mechanism is sufficient.

2.4.4 Dependence of VA bias on contrast

Effect of contrast on type II plaids - Yo and Wilson (1992)

Phenomena: Yo and Wilson (1992) reported that the bias towards VA in type II plaids consistently increased with reduced contrast. For example, Figure 2-24a,c show a type II plaid at high contrast and at low contrast. For durations over 100msec the high contrast plaid is perceived as moving in the veridical direction, while the low contrast is heavily biased towards the VA solution (Yo and Wilson, 1992).

Model Results: Figure 2-24b,d show the VA, IOC and Bayesian predictions for this stimulus. Obviously, both VA and IOC solutions are unaffected by the contrast

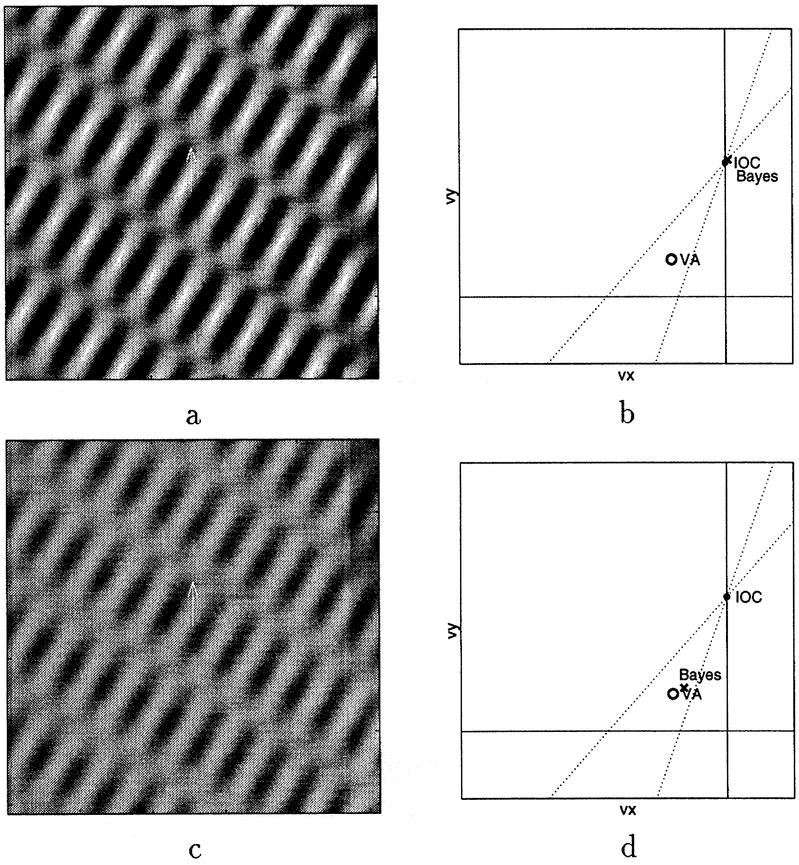


Figure 2-24: A high contrast type II plaid (a) viewed at long durations, may be perceived veridically, but the same stimulus at low contrast (b) shows a strong VA bias (Yo and Wilson 92). As shown in (b) and (d) the VA and IOC predictions are not affected by contrast, but the Bayesian estimator with a fixed σ shows the same shift from veridical to biased as contrast is decreased.

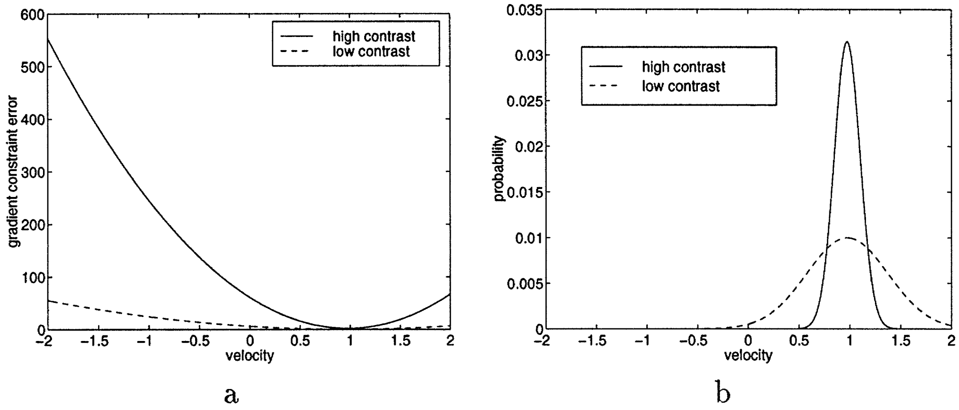


Figure 2-25: **a.** The local consistency (equation 2.5) for various vertical velocities measured at a single location in the stimulus shown in figure 2-24. At low contrast there is only a small difference between the degree to which the true velocity satisfies the gradient constraint and the degree to which other velocities do so. **b.** The local likelihood (equation 2.6) for various vertical velocities at the same location. At low contrast there is a higher degree of uncertainty.

and hence cannot by themselves account for the percept. The Bayesian estimate, on the other hand, changes from veridical to biased as contrast is decreased even though the only free parameter σ is held constant.

Discussion: To gain intuitive understanding of the change in the Bayesian prediction as contrast as varied, recall from section 2.3.1 that the contrast changes the "fuzziness" of the constraint line. Thus at low contrast, both constraint lines are very fuzzy, and the VA solution receives relatively high likelihood relative to the IOC solution. We emphasize that this change in "fuzziness" with contrast does not have to be put in especially to explain this phenomena. It is a direct consequence of the probabilistic formulation – at low contrast there is more uncertainty locally. Figure 2-25a shows the consistency measure (equation 2.5) for different vertical velocities measured at a single location in the stimulus shown in figure 2-24. At low contrast there is only a small difference between the degree to which the true velocity satisfies the gradient constraint and the degree to which other velocities do so. Therefore when the local likelihoods are calculated (equation 2.6) one obtains figure 2-25b. At lower contrast the likelihood function is less peaked, and there is more local uncertainty.

While the sharpness of the local likelihoods change with contrast, the prior proba-

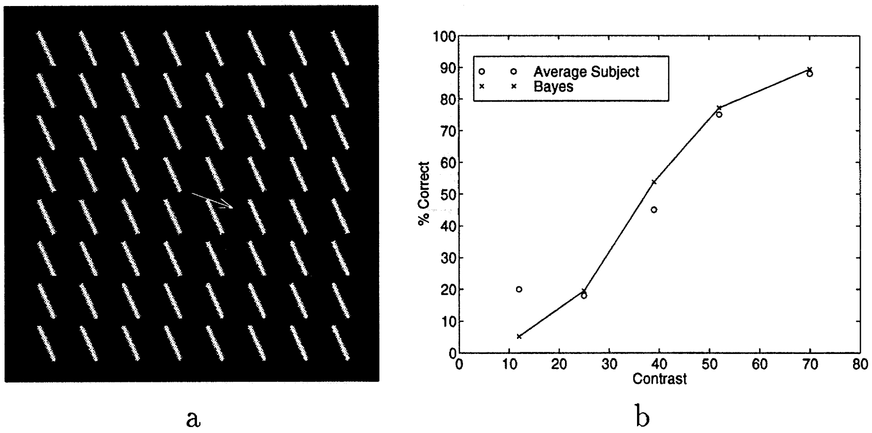


Figure 2-26: A stimulus used by Lorenceau et al. (93) suggesting the need for independent terminator and line motion mechanisms. A matrix of lines moves oblique to the line orientations. At high contrast the motion of the lines is veridical while at low contrast it is misperceived **a.** A single frame from the sequence. **b.** The results of a two alternative forced choice experiment (up/down) replotted from Lorenceau et al. (1992) (average subject shown with circles). The solid line shows the predictions of the Bayesian model. A single Bayesian mechanism would predict systematic errors at low contrast with an increase in correct responses as contrast is increased.

bility does not change. As mentioned earlier, the prior probability of the VA solution is higher, and hence at low contrasts the Bayesian solution is biased towards the VA. At high contrast, however, as the likelihoods become much more peaked, the prior has less influence and the Bayesian estimate approaches the IOC solution.

2.4.5 Contrast effects on line stimuli - Lorenceau et al 1992

Phenomena: Lorenceau et al. (1993) asked subjects to judge whether a matrix of oriented lines moved above or below the horizontal (see figure 2-26a) as the contrast of the display was systematically varied. The results are replotted in figure 2-26b. Note that at low contrasts, performance is far below chance indicating subjects perceived upward motion while the patterns moved downward. Lorenceau et al. modeled these results using two separate mechanisms, one dealing with terminator and other with line motion. The terminator mechanism is assumed to be active primarily at high contrast and the line mechanism at low contrast.

Model Results: The solid line in figure 2-26b shows the simulated performance of the Bayesian model on this task. Again, the percentage of correct responses is obtained by using a "soft" threshold on the model's predicted direction of motion. Although the model does not include separate "terminator" and "line" motion mechanisms, it predicts a gradual shift from downward motion to upward motion as contrast is increased. The parameter σ is held fixed.

Discussion: The intuition behind the model's performance in this task is similar to the one in the plaid displays. At high contrast, the likelihood is peaked and the estimated motion is veridical. At low contrast, however, the likelihood at the endpoints of the lines and along the lines, is more "fuzzy" and the prior favoring slow velocities has a large influence. Hence, motion is perceived in the normal velocity which is slower than the veridical one. There is no need to assume separate terminator and line mechanisms.

Influence of contrast on the speed of a single grating - Thompson et al 1996

Phenomena: Thompson et al. (1996) have shown that the perceived speed of a single grating depends on the contrast. Noting that "lower-contrast patterns consistently appear to move slower", they conducted an experiment in which subjects viewed a high contrast (70%) grating followed by a test low contrast (10%) grating. The subjects adjusted the speed of the test grating until the perceived speeds were matched (see figure 2-27a). Although the magnitude of the effect varied slightly between subjects, the direction of the effect was quite robust. Typical results are shown in figure 2-27b. In order to match the perceived speed of the low contrast grating, the high contrast grating needs to move about 70% slower. Similarly, in order to match the perceived speed of the high contrast grating, the low contrast grating needs to move about 150% faster.

Model Results: Figure 2-27c shows the output of a Bayesian estimator on this stimulus. For a fixed σ the low contrast grating is predicted to move slower. The predicted speed match is computed by dividing the estimated speeds of the two gratings.

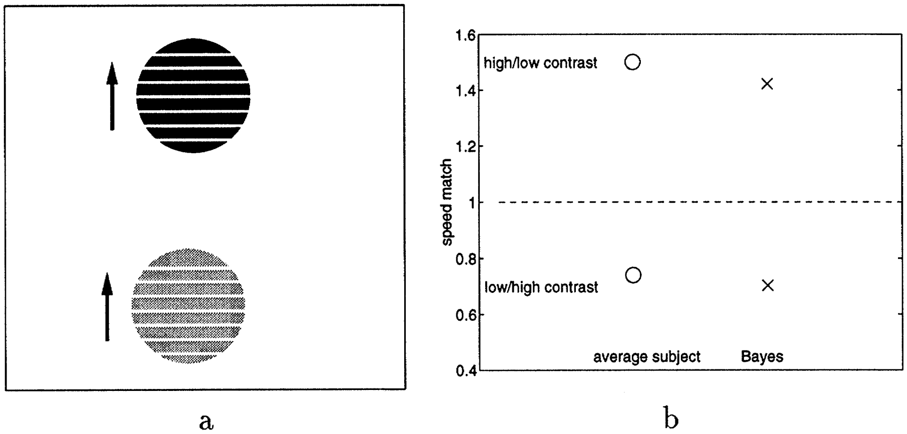


Figure 2-27: An experiment conducted by Thompson et al. (1996) showing that low contrast stimuli appear to move slower. Subjects viewed a high contrast grating (70%) followed by a test low contrast grating (10%). They adjusted the speed of the test grating until the perceived speeds were matched. **b.** Circles show the results averaged over 6 subjects replotted from (Thompson et al., 1996). In order to match the perceived speed of a low contrast grating, the high contrast grating needs to move about 70% slower. Similarly, in order to match the perceived speed of a high contrast grating, the low contrast grating needs to move about 150% faster. Crosses show the output of the Bayesian estimator. At low contrast, the likelihood is less peaked and the prior favoring slow speeds dominates. Hence the low contrast grating is predicted to move slower.

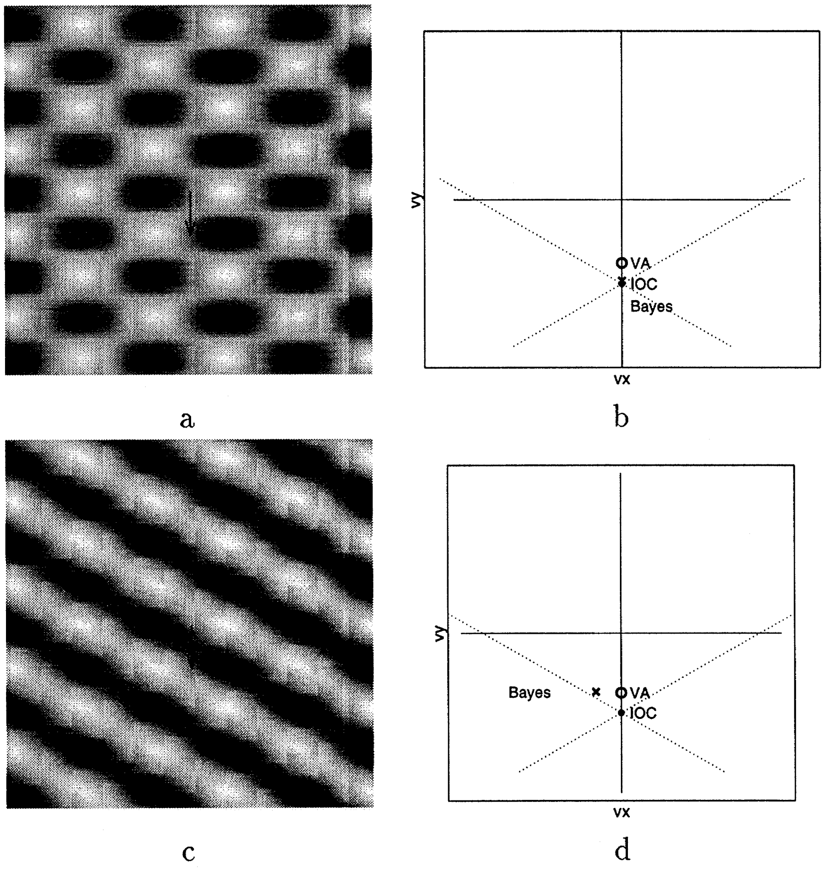


Figure 2-28: The influence of relative contrast on the perceived direction of a moving type I plaid (Stone et al., 1990). When both components are of identical contrasts the perceived motion is in the veridical direction. When they are of unequal contrasts, the perceived direction is biased in the direction of the higher contrast grating. A similar pattern is observed in the output of the Bayesian estimator.

Discussion: Again, at at low contrast the likelihood is less peaked and the prior favoring slow speeds dominates. Hence the low contrast grating is predicted to move slower than a high contrast grating moving at the same speed.

Dependence of type I direction on relative contrast - Stone et al. (1990)

Phenomena: Stone et al. (1990) showed subjects a set of type I plaids and varied the ratio of the contrasts between the two components. They found that the direction of motion of the plaid was biased in the direction of the higher contrast grating. The magnitude of the bias changed as a function of the "total contrast" of the plaid, i.e. the sum of the contrasts of the two gratings. When the contrast of both gratings was

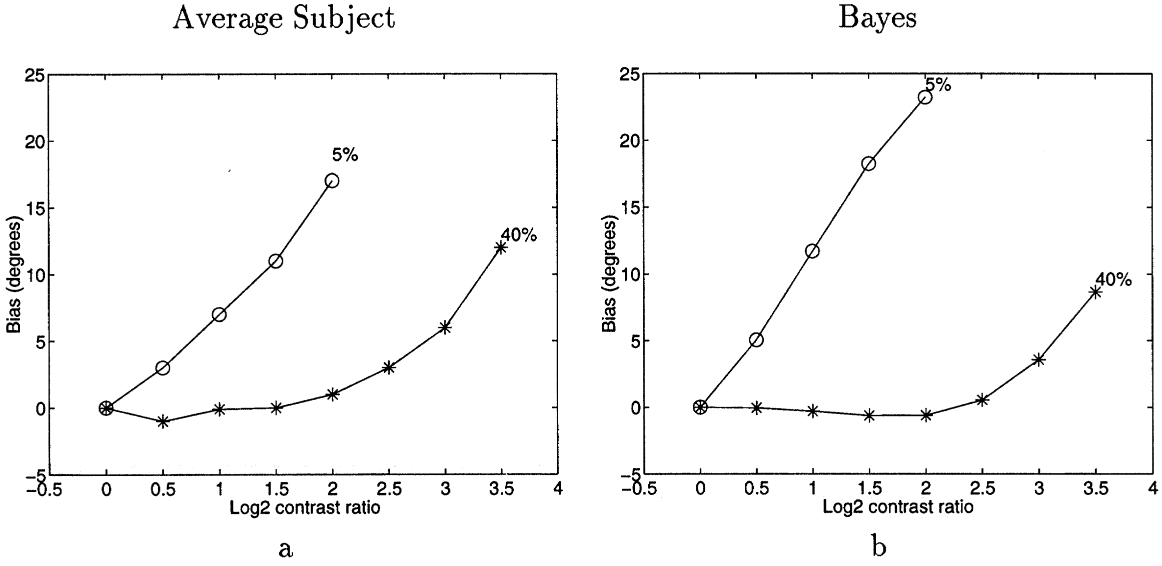


Figure 2-29: An experiment conducted by Stone et al. (1990) showing the influence of relative contrast on the perceived direction of a moving plaid. Subjects viewed a set of type I plaids and the contrasts of the two components was systematically varied. **a.** Results averaged over subjects replotted from. (Stone et al., 1990). The direction of motion of the plaid was biased in the direction of the higher contrast grating and the magnitude of the bias decreases with increased total contrast. **b.** The Bayesian estimator gives similar results. (cf. (Heeger and Simoncelli, 1991)).

increased (while the ratio of contrast stayed constant) a smaller bias was observed. Figure 2-29a shows data averaged over subjects replotted from (Stone et al., 1990).

Model Results: The results of the Bayesian estimator are shown in figure 2-29b. Similar to the results of human observers, the estimate is biased in the direction of the higher contrast grating and the magnitude of the bias decreases with increasing total contrast.

Discussion: Again this is a result of the fact that as contrast is decreased the local uncertainty decreases. Thus in figure 2-28d, the likelihood corresponding to the low contrast grating is a very "fuzzy" constraint line. In this case, although the Bayesian solution does not lie exactly on both constraint lines it has very similar likelihood to the IOC solution. In terms of the prior, however, the Bayesian solution is favored because it is slower. When both gratings are of identical contrasts, the likelihoods have equal fuzziness and the Bayesian solution has the correct direction (although the magnitude is smaller than the IOC solution). When the total contrast is increased, all

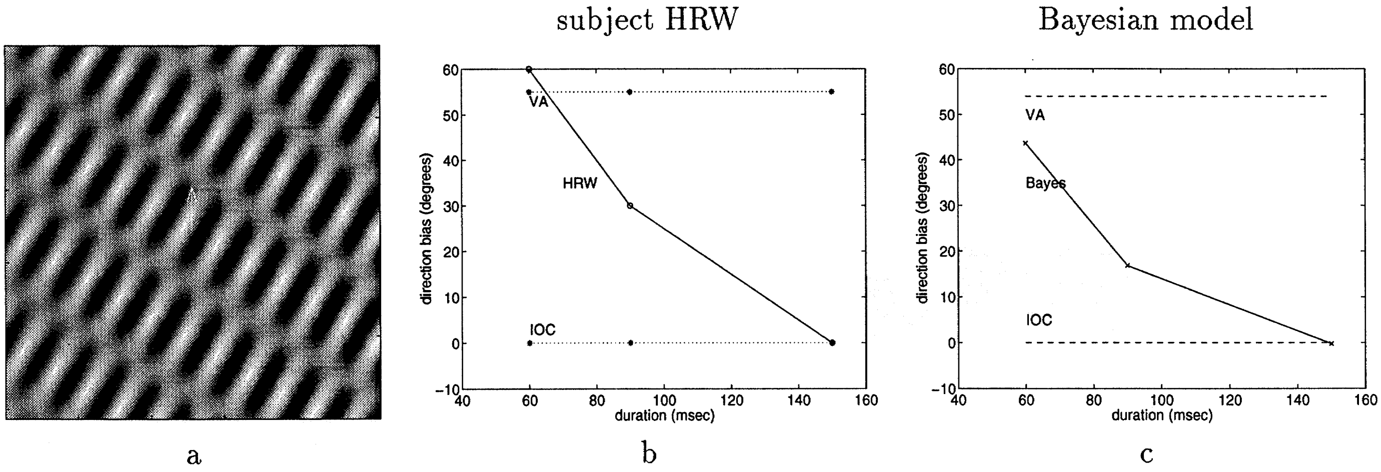


Figure 2-30: The influence of duration on performance in the experiment conducted by Yo and Wilson (1992). At short durations, the perceived motion is heavily biased towards the VA, and it approaches the IOC solutions at long durations. **a.** a single frame from the sequence. **b.** The results of subject HRW replotted from (Yo and Wilson, 1992). **c.** The predictions of a Bayesian estimator. The predicted velocity shows a gradual shift from VA to IOC as duration increases.

the likelihoods become more peaked and the Bayesian solution is forced to lie closer to the IOC solution.

Although the results of the Bayesian estimator is in qualitative agreement with the psychophysical results for this task, the quantitative fit can be improved. Heeger and Simoncelli (1991) have obtained better fits for this data using their model that also includes a nonlinear gain control mechanism.

2.4.6 Dependence of bias on duration

Dependence of type II bias on duration - Yo and Wilson (1992)

Phenomena: Yo and Wilson (1992) reported that the perceived direction of type II plaids changes with stimulus duration. At short durations, the perceived direction is heavily biased in the direction of the vector average and gradually approaches the IOC solution as duration is increased. Figure 2-30b shows the results of a single subject.

Model Results: Figure 2-30c shows the predictions of the Bayesian estimator. The model was given 5 frames of the video sequence, and the local likelihood was calculated

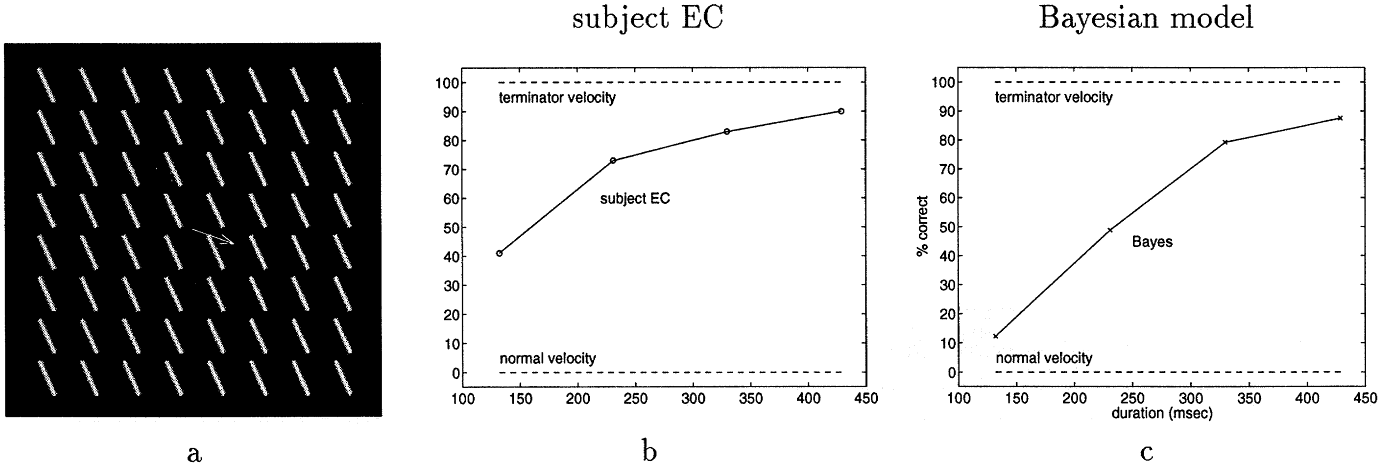


Figure 2-31: The influence on duration on performance in the experiment conducted by Lorenceau et al. (93). At short duration, performance is below chance indicating subjects perceive motion in the normal direction, while at long durations the perceived motion is largely veridical. **a.** a single frame from the sequence. **b.** The results of a single subject replotted from (Lorenceau et al., 1992). Despite significant individual variations, subjects consistently perform below chance at short durations and improve as duration increases. **c.** The predictions of a Bayesian estimator. A single Bayesian mechanism would predict systematic errors at short durations with an increase in correct responses as duration is increased.

by summing filter outputs over space and time. In that respect the results in this section differ from those reported in other sections, where only two frames were used to calculate the local likelihoods. Note the change in model output with increased duration.

Discussion: As discussed in section 2.3.1, short durations serve to make the local likelihood less peaked. In fact, the short duration acts in the model much like low contrast (figure 2-25). At short durations, there is only a small difference between the degree to which the true velocity satisfies the gradient constraint and the degree to which other velocities do so. However, as gradient information is combined over time, the difference becomes more pronounced and the uncertainty in the local measurement decreases. The shorter the presentation time the more the local information is ambiguous.

While the sharpness of the local likelihood change with duration, the prior probability does not. Hence the VA solution which has a higher prior probability is favored at short durations, while at long durations the Bayesian estimate approaches the IOC solution.

Dependence on duration in line drawings – Lorenceau et al. 1992

Phenomena: Lorenceau et al. (1992) reported a similar effect of duration in the discrimination of line motion. As explained in the previous section, subjects were requested to judge whether the matrix of lines moved above or below the horizontal. At short durations, they found that performance was below chance, indicating that subjects perceived the lines moving in the normal direction, but performance improved at longer durations. Figure 2-31b shows the results of a single subject replotted from (Lorenceau et al., 1992). Despite significant individual variations, subjects consistently perform below chance at short durations and improve as duration increases.

Model Results: Figure 2-31c shows the output of the Bayesian estimator. A single mechanism predicts systematic errors at short durations with an increase in correct responses as duration is increased. Note that this explanation does not require separate "1D" and "terminator" mechanisms. Rather it is explained in the same way as the influence of duration on plaids.

Again, at low durations all local measurements have higher degree of uncertainty. In the Bayesian model there is no categorization of location into "1D" or "2D" but at all locations the gradient constraint is accumulated over space and time. At short durations, therefore, there is less signal in the local spatiotemporal window, and hence more uncertainty in the local likelihoods. In this condition, the prior favoring slow speeds dominates and perception is in the normal direction. At long durations, the local uncertainty is decreased, and the prior has a much weaker influence.

Discussion: The results reported in this section were obtained by using a spatiotemporal Gaussian window in equation 2.6. This gives an additional free parameter to fit the data. However the qualitative nature of the results are unchanged when the window function is changed. Any summation of information over time would lead to a decrease in local uncertainty with longer durations. Thus a Bayesian estimation strategy predicts highly biased estimates at low durations but more veridical velocity as duration increases.

2.4.7 Non-translational motions

So far we have discussed stimuli undergoing uniform translation. Although the model returns a flow field we could capture it with a single velocity vector. Now we show the output of the model on non-translational motions. We display the output of the model by plotting arrows at different (arbitrarily chosen) locations of the image.

Circles and derived figures in rotation - Wallach 1956

Phenomena: Musatti (1924) and Wallach et al. (1956) observed that when circular figures are rotated in the image plane (e.g by putting them on a turntable) they are not perceived as rigidly rotating. A rotating circle appears static, a rotating spiral appears to contract, and a rotating ellipse appears to deform nonrigidly. In the case of the rotating ellipse, Wallach et al. (1956) noted that the perceived rigidity is most pronounced when the ellipse is "fat" — with aspect ratio close to unity. Musatti pointed out that when a small number of rotating features are added to the display, the rigid percept becomes prominent.

Model Results: Figure 2-32 shows the output of the Bayesian estimator on these stimuli. As in human perception the rotating circle is perceived as static, the rotating spiral as expanding and the rotating ellipse as deforming nonrigidly. Figure 2-33 shows the model output on a narrow ellipse and on an ellipse with four rotating features added. Note that in this case, consistent with human perception, the predicted motion is much closer to rotation. The parameter σ is held constant.

Discussion: Why does the model "misperceive" these motions? First note that for the stimuli in figure 2-32, the perceived motions and the rotational motions have very similar likelihoods. That is, due to the low curvature of the figure, the local likelihoods are highly ambiguous. Given that the likelihoods are nearly identical, the Bayesian estimator is dominated by the prior. Here again, the "slowness" prior may be responsible for the percept. Figure 2-34 shows the total magnitude of the velocity

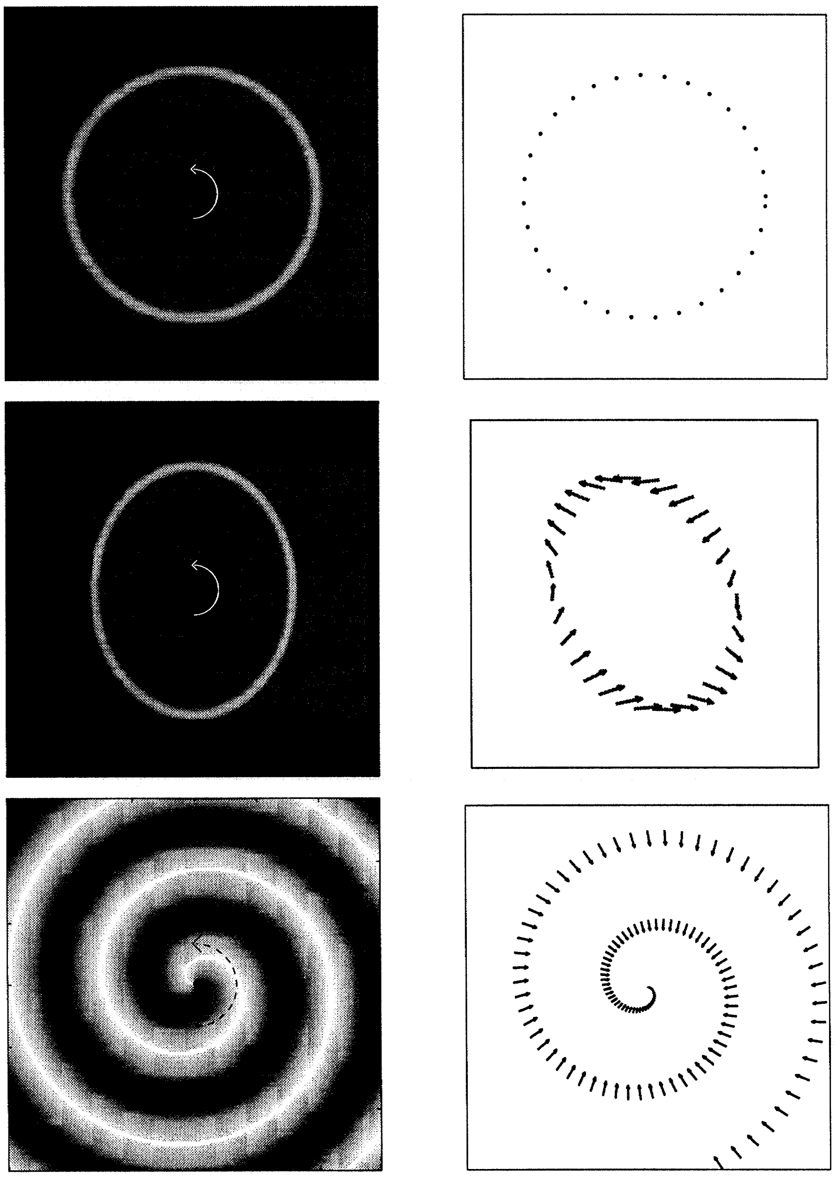


Figure 2-32: Biased perception in Bayesian estimation of circles and derived figures in rotations. Due to the prior favoring slow and smooth velocities, the estimate may be biased away from the veridical velocity and towards the normal components. These biases are illustrated here. A rotating circle appears to be stationary, a rotating ellipse appears to deform nonrigidly, and a rotating spiral appears to expand and contract.

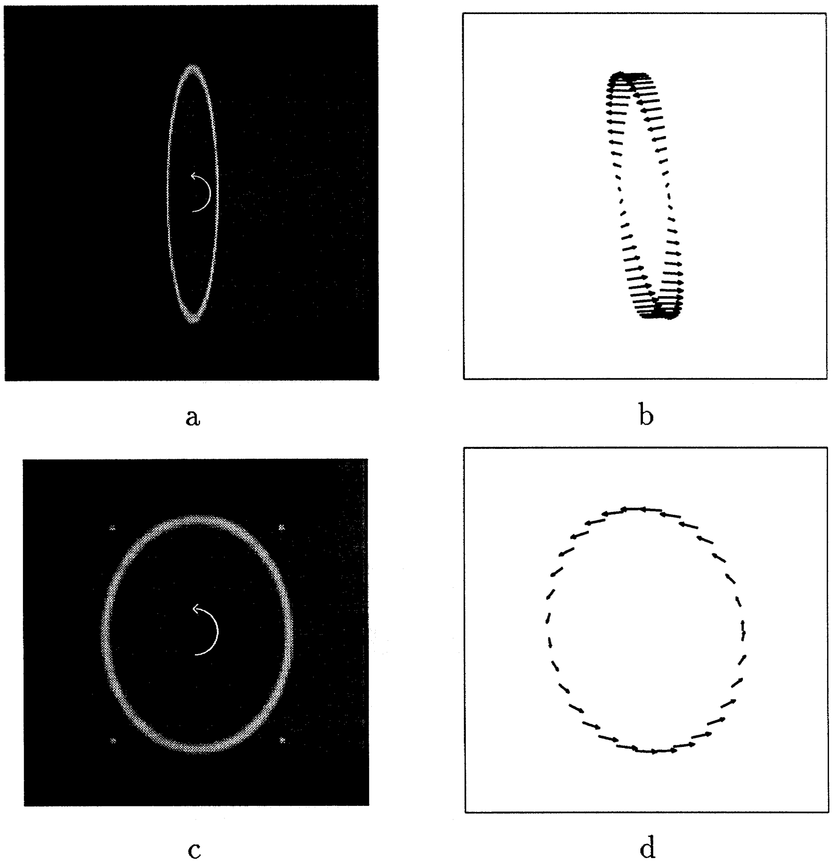


Figure 2-33: The percept of nonrigid deformation is influenced by stimulus shape and by additional features. For a "narrow" rotating ellipse, the Bayesian estimate is similar to rotation. Similarly, for a "fat" rotating ellipse with four rotating dots, the estimate is similar to rotation. This is consistent with human perception. The parameter σ is held constant.

fields. Note that the rotational velocity is much faster than the Bayesian estimate, and hence is not favored.

The Bayesian estimate considers both the likelihood and the prior. Thus once the rotating stimulus includes locations that are relatively ambiguous (e.g. the endpoints of a narrow ellipse, or dots flanking the fat ellipse), the estimate resembles rotation. The rotation still has lower prior probability but high likelihood.

A slightly different account of these illusions was given by Hildreth (1983). Her model chooses the velocity field of least variation that satisfies the gradient constraint at every location along the ellipse. Although her algorithm did not include an explicit penalty for fast velocity fields it gave similar results to those shown here – a rotating circle was estimated to be stationary, a rotating spiral was estimated to be expanding and a rotating fat ellipse was estimated to be deforming.

Note however that by penalizing the magnitude of the first derivative, Hildreth's algorithm includes an implicit penalty for fast non-translational velocity fields. That is, for all translational velocity fields, the first derivative is zero everywhere and there is no distinction between fast and slow fields. For velocity fields whose first derivative does not vanish, however, the magnitude of the first derivative increases with increased speed. Thus Hildreth's algorithm will in general prefer a slow deformation to a faster rotation. It will not, however, prefer a slow translation to a faster one, and thus can not account for biases encountered in translating stimuli (e.g. the VA bias in plaids).

Smooth curves in translation - Nakayama and Silverman 1988

Phenomena: Nakayama and Silverman (1988) found that smooth curves including sinusoids, Gaussians and sigmoids, may be perceived to deform nonrigidly when they are translated rigidly in the image plane. Figure 2-35a shows an example. A "shallow" sinusoid is translating rigidly horizontally. This stimulus is typically perceived as deforming nonrigidly. The authors noted that the perceived nonrigidity was most pronounced for "shallow" sinusoids in which the curvature of the curves was small.

Model Results: Figure 2-35b shows the output of the Bayesian estimator. For the shallow sinusoid the Bayesian estimator favors a slower hypothesis than the veridical

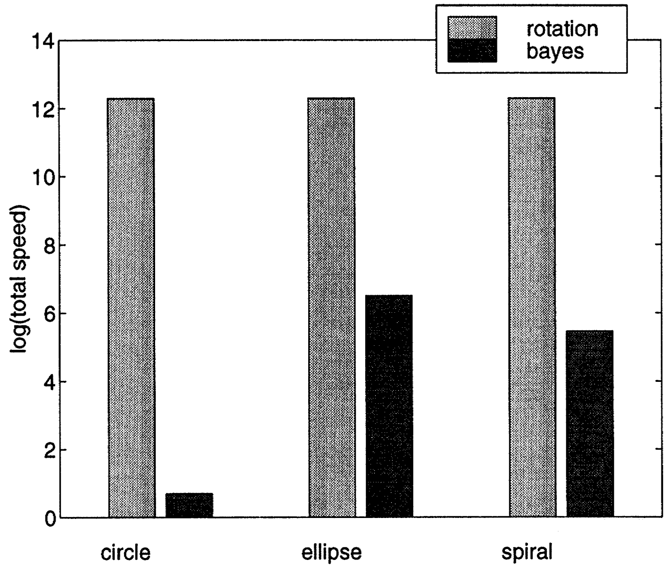


Figure 2-34: The total magnitude of the velocity fields arrived at by the Bayesian estimate for the stimuli in 2-32 as compared to the true rotation. Note that the rotational velocity is much faster than the Bayesian estimate, and hence is not favored.

rigid translation. Figure 2-35d shows the output on the sharp sinusoid. Note that a fixed σ gives a nonrigid percept for the shallow sinusoid and a rigid percept for the sharp sinusoid.

Discussion: Again this is the result of the tradeoff between "slow" and "smooth" priors. The nonrigid percept is slower than the rigid translation but less smooth. For shallow sinusoids, the nonrigid percept is still relatively smooth, but for sharp sinusoids the smoothness term causes the rigid percept to be preferred. The shape of sinusoid for which the percept will shift from rigid to nonrigid depends on the free parameter λ which governs the tradeoff between the slowness and smoothness terms. The qualitative results however remain the same — sharp sinusoids are perceived as more rigid than shallow ones. Similar results were also obtained with the other smooth curves studied by Nakayama and Silverman — the Gaussian and the sigmoidal curves.

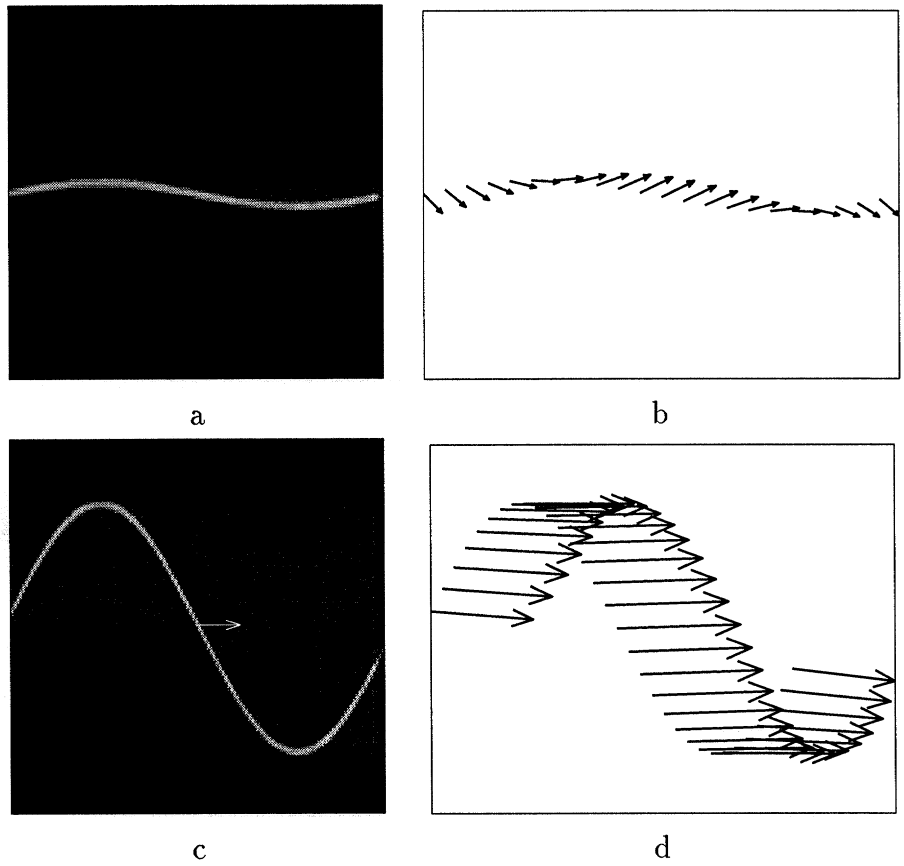


Figure 2-35: **a.** A "shallow" sinusoid translating horizontally appears to to deform nonrigidly (Nakayama and Silverman 1988). **b.** The nonrigid deformation is also prevalent in the Bayesian estimator. **c.** A "sharp" sinusoid translating horizontally appears to translate rigidly (Nakayama and Silverman 1988). **d.** Rigid translation is also prevalent in the Bayesian RBF estimator.

2.5 Discussion

Since the visual system receives information that is ambiguous and uncertain, it must combine the input with prior constraints to achieve reliable estimates. A Bayesian estimator is the simplest reasonable approach and the prior favoring slow and smooth motions offer reasonable constraints. In this paper we have asked how such a system will behave. We find that, like humans, its motion estimates include apparent biases and illusions. Moreover, this non-veridical perception is quite similar to that exhibited by humans in the same circumstances.

In recent years a large number of phenomena have been described in velocity estimation, usually connected with the aperture problem. In reviewing a long list of phenomena, we find that the Bayesian estimator almost always predicts the psychophysical results. The predictions agree qualitatively, and are often in remarkable agreement quantitatively.

The Bayesian estimator is a simple and reasonable starting point for a model of motion perception. Insofar as it explains the data, there is no need to propose specific mechanisms that deal with lines, terminators, plaids, blobs etc. These other mechanisms are often poorly defined, and they are often assumed to turn on or off according to special rules.

The Bayesian estimator described here can be applied to any image sequence that contains a single moving surface. It works with gratings, plaids, ellipses or spirals without modification. It usually needs only a single free parameter σ, which corresponds to the noise or internal uncertainty level in the observer's visual system. Even this parameter remains fixed when the individual observer and viewing conditions are fixed.

Beyond the specifics of our particular model, we have shown that human motion perception exhibits two fundamental properties of a Bayesian estimator. First, observers give different amounts of weight to information at different locations in the image - e.g. a small number of features can profoundly influence the percept and high contrast locations have greater influence than low contrast ones. This is consistent with a Bayesian mechanism that combines sources of evidence in accordance with their uncertainty. Second, the motion percept exhibits a bias towards slow and smooth velocities, consistent with a Bayesian mechanism that incorporates prior knowledge as well as evidence into the estimation.

Each of these properties have appeared in some form in previous models. The notion of giving unequal weight to different motion measurements appears, for example, in the model suggested by Lourenceau et al. (1992). Mingolla et al. (1992) suggested assigning these weights according to their "saliencies" which would in turn depend on contrast. In the Bayesian framework, the amount of weight given to a particular measurement has a concrete source — it depends on its uncertainty. Thus the low weight given to low contrast, short duration or peripherally viewed features is a consequence of the high degree of uncertainty associated with them. Moreover, there is no need to arbitrarily distinguish between "2D" and "1D" local features — all image regions have varying degrees of uncertainty, and the strong influence of cornerlike features is a consequence of the relatively unambiguous motion signals they give rise to.

As mentioned in the introduction, the models of Hildreth (1983) and Grzywacz and Yuille (1991) include a bias towards smooth velocity fields. However these algorithms do not have the concept of varying degrees of ambiguity in local motion measurements. They either represents the local information as a constraint line in velocity space or as a completely unambiguous 2D measurement. They therefore can not account for the gradual shift in perceived direction of figures as contrast and duration are varied.

The smoothness assumption used by Hildreth (1983) and others, can be considered a special case of the regularization approach to computational vision introduced by Poggio et al. (1985). This approach is built on the observation that many problems in vision are "ill-posed" in the mathematical sense — there are not enough constraints in the data to reliably estimate the solution. Regularization theory (Tikhonov and Arsenin, 1977) provides a general mathematical framework for solving such ill-posed problems by minimzing cost functions that are the sum of two terms – a "data" term and a "regularizer" term. For many problems, Bayesian MAP estimation and regularization theory give mathematically equivalent algorithms (e.g. (Marroquin et al., 1987)). This is also true for the motion theory presented here — although we have used the language of Bayesian inference we could have equivalently described the theory in terms of regularization. In the appendix we make this mapping explicit.

The model of Heeger and Simoncelli (1991) was to the best of our knowledge, the first to provide a Bayesian account of human motion perception that incorporatea a prior favoring slow speeds. Indeed the first stage of our model, the extraction of local likelihoods, is very similar to the Heeger and Simoncelli model. In our model, however, these local likelihoods are then combined across space to estimate a spatially varying velocity field. In spatially isotropic stimuli (such as plaids and gratings) there is no need to combine across space as all spatial locations give the same information. However, integration across space is crucial in order to account for motion perception in more general stimuli such as translating rhombuses, rotating spirals or translating sinusoids.

Another local motion analysis model was introuced by Bulthoff et al. (1989) who described a simple, parallel algorithm that computes optical flow by summing activities over a small neighborhood of the image. Unlike the Heeger and Simoncelli model, their model did not include a prior favoring slow velocities and therefore predicts the IOC solution for all plaid stimuli.

We have attempted to make the Bayesian estimator discussed here as simple as possible, at the sacrifice of biological faithfulness. Thus we assume a Gaussian noise model, a fixed σ and linear gradient filters. One disadvantage of this simple model is that in order to obtain quantitative fits to the results of existing experiments we had to vary σ between experiments (but σ was always held fixed when modeling a single experiment with multiple conditions). Although changing σ does not in general change the qualitative nature of the Bayesian estimate, it does change the quantitative results. A more complicated Bayesian estimator, that also models the nonlinearities in early vision, may be able to fit more data with fixed parameters.

How could a Bayesian estimator of the type discussed here be implemented given what is known about the functional architecture of the primate visual system? The local likelihoods are simple functions (squaring and summing) of the outputs of spatiotemporal filters at a particular location. Thus a population of units in primary visual cortex may be capable of representing these local likelihoods (Heeger and Simoncelli, 1991). Combining the likelihoods and finding the most probable velocity estimate, however, is a more complicated matter and is an intriguing question for future research.

Indeed understanding the mechanism by which human vision combines local motion signals may prove fruitful in the design of artificial vision systems. Human motion perception seems to accurately represent uncertainty of local measurements, and to combine these measurements in accordance with their uncertainty together with a prior probability. Despite this sophistication motion perception is immediate and effortless, suggesting that the human visual system has found a way to perform fast Bayesian inference.

2.6 Appendix

2.6.1 Solving for the most probable velocity field

We derive here the equations for finding the parametric vector that maximizes the posterior probability. To simplify the notation, we denote the location (x, y) with a single vector r. Assume that the velocity field $v(r)$ is composed of a sum of N basis functions with the coefficients defined by the parameter vector θ. Define $\Psi(r)$ a 2 by N matrix which give the two components of the basis functions at location r, then $v(r) = \Psi(r)\theta$. Using this notation we can now rewrite the likelihoods and the prior as a function of θ.

Recall that the local likelihood is given by:

$$L_r(v) = \alpha e^{-\sum_r w(r)(I_x v_x + I_y v_y + I_t)^2/2\sigma^2} \tag{2.14}$$

(we use the convention that for any probability distribution α represents the normalization constant that guarantees that the distribution sum to unity). By completing the square, this can be rewritten:

$$L_r(v) = \alpha e^{-(v-\mu(r))^t \Sigma^{-1}(r)(v-\mu(r))/2\sigma^2} \tag{2.15}$$

where $\mu(r), \Sigma^{-1}(r)$ represent the mean and covariance matrices of the local likelihood.

$$\Sigma^{-1}(r) = \sum_s w_{rs} \begin{pmatrix} I_x^2(s) & I_x(s)I_y(s) \\ I_x(s)I_y(s) & I_y^2(s) \end{pmatrix} \tag{2.16}$$

and $\mu(r)$ a solution to:

$$\Sigma^{-1}(r)\mu(r) = y(r) \tag{2.17}$$

with

$$y(r) = \sum_s w_{rs} \begin{pmatrix} I_x(s)I_t(s) \\ I_y(s)I_t(s) \end{pmatrix} \tag{2.18}$$

Substituting $v(r) = \Psi(r)\theta$ into equation 2.15 gives the local likelihood of the image derivatives given θ:

$$L_r(\theta) = \alpha e^{-(\Psi(r)\theta-\mu(r))^t \Sigma^{-1}(r)(\Psi(r)\theta-\mu(r))/2\sigma^2} \tag{2.19}$$

and finally assuming conditional independence, the global likelihood for the image derivatives is given the product of the local likelihoods at all locations:

$$L(\theta) = \Pi_r L_r(\theta) \tag{2.20}$$

We now express the prior probability as a function of θ. Recall that the prior favors slow and smooth velocities:

$$P(V) = \alpha e^{-\sum_r (Dv)^t(r)(Dv)(r))/2} \tag{2.21}$$

where D is a differential operator. Substituting $v(r) = \Psi(r)\theta$ gives the prior probability on θ:

$$P(\theta) = \alpha e^{-\theta^t R\theta/2} \tag{2.22}$$

Where R is a symmetric, NxN matrix such that

$$R_{ij} = \sum_r (D\Psi_i^t)(r)(D\Psi_j)(r) \tag{2.23}$$

where we have used $\Psi_i(r)$ the ith basis field, and $D\Psi_i(r)$ the results of applying the differential operator D to that basis field.

The posterior is given by:

$$P(\theta|I) = \alpha P(\theta)P(I|\theta) \tag{2.24}$$

The log-posterior is given by:

$$\begin{aligned} \log P(\theta|I) = k - \theta^t R\theta/2\sigma_p^2 \\ + \sum_r -(\Psi(r)\theta - \mu(r))^t \Sigma^{-1}(r)(\Psi(r)\theta - \mu(r))/2\sigma^2 \end{aligned} \tag{2.25}$$

(note that the log-posterior is quadratic in θ or in other words the posterior is a Gaussian distribution. Thus maximizing the posterior is equivalent to taking its mean)

To find θ^* the value of θ that maximizes the posterior we solve:

$$A\theta^* = b \tag{2.26}$$

with:

$$A = \left(\sum_r \Psi^t(r)\Sigma^{-1}(r)\Psi(r)/\sigma^2 + R/\sigma_p^2\right) \tag{2.27}$$

$$b = \left(\sum_r \Psi^t(r)\Sigma^{-1}\mu(r)\right)/\sigma^2 \tag{2.28}$$

Specifically, the parameters we use in these simulations are as follows. The differential operator D was chosen so that the Green's functions corresponding to it were Gaussians with standard deviation equal to 70% of the size of the image. The basis fields were also Gaussians with the same standard deviations. We used 50 basis fields, 25 with purely horizontal velocity and 25 with pure vertical velocity. The centers of the basis fields were equally spaced in the image, i.e. were placed on a $5x5$ grid. In this case the matrix R has a particularly simple form. If Ψ_i and Ψ_j are both vertical (or horizontal) then R_{ij} is simply the value of the ith basis field evaluated at the center of the jth basis field. Otherwise, $R_{ij} = 0$.

To summarize, given an image sequence and a parameterization of the velocity field, the Bayesian estimate of motion is obtained by solving equation 2.26. Finally the optimal velocity field is obtained by $v(r) = \Psi(r)\theta^*$.

2.6.2 Relation to regularization theory

There are very close links between Bayesian MAP estimation and regularization theory (e.g. (Marroquin et al., 1987)). For completeness, we now show how to rephrase the Bayesian motion theory presented here in terms of regularization theory.

Regularization theory calls for minimizing cost functions that have two terms: a "data" term and a "regularizer term". A classical example is function approximation where one is given samples $\{x_i, y_i\}$ and wishes to find the approximating function. Obviously this is an ill-posed problem – there are an infinite number of functions that could approximate the data equally well. A typical regularization approach calls for minimizing:

$$J(f) = \sum_i (f(x_i) - y_i)^2 + \lambda \int_x \|Df(x)\|^2 dx \tag{2.29}$$

The first term on the right hand side is the data term and the second term is the regularizer, in this case regularization is performed by penalizing for high derivatives.

Note that the log posterior in equation 2.25 can also be decomposed into two terms that depend on θ . The sum of the log likelihoods $\sum_r -(\Psi(r)\theta - \mu(r))^t \Sigma^{-1}(r)(\Psi(r)\theta - \mu(r))/2\sigma^2$ and the log prior $-\theta^t R\theta/2\sigma_p^2$. In the language of regularization theory, the negative sum of the log likelihoods would be the "data term" and the negative log posterior would be the "regularizer term".

The negative log posterior, when considered as a "regularizer" is quite similar to the smoothness regularizer in equation 2.29 in that it penalizes for values of θ that correspond to velocity fields that have large derivatives. Likewise the negative log likelihood is similar to the data term in equation 2.29 in that it penalizes for the squared error between the observed data and the predicted velocity field. The main difference, however, is that different observations are given different weights in the log posterior. Recall from section 2.2 that in Bayesian MAP estimation for Gaussian likelihoods the weight of an observation is inversely proportional to its variance, hence the Σ^{-1} factor in equation 2.25. Although the regularization framework is broad enough to encompass nonuniform weights for the data, the use of uniform weighting as in equation 2.29 is most common (but see (Girosi et al., 1990) for an exception).

An elegant result that can be derived in the regularization framework shows that the function f that minimizes J in equation 2.29 can be expressed as a superposition of basis functions (see (Girosi et al., 1995) and references within). In contrast, here we assume a particular representation for the velocity field rather than deriving it. We do this because the number of basis functions required for the optimal function

f is equal to the number of datapoints. In the case of motion analysis, this number is prohibitively large and for computational efficiency we prefer a low dimensional representation. A similar approach has been used in function approximation and time series prediction (Broomhead and Lowe, 1988; Poggio and Girosi, 1989) when the number of datapoints is large. We have found that as long as one uses the prior over velocity fields, the exact form of the representation used is not crucial — very similar results are obtained with different representations (Weiss, 1997).

Chapter 3

Smoothness in Layers – a framework for motion estimation and segmentation in human vision

Abstract

A large body of experimental results in human motion perception is consistent with the notion that the visual system pools motion information across space by assuming that nearby points in the image have similar velocities. Although such "smoothness" based models successfully account for the percept in scenes containing a single motion, they can not by themselves account for percepts in scenes containing multiple motions. More elaborate models that allow discontinuities in the motion field, or that restrict smoothness to contours are inconsistent with human perception.

In this paper we suggest an extension of the smoothness assumption to scenes containing multiple motions. We assume the scene contains a small number of surfaces or layers, and that velocity varies smoothly within a layer but not across layers. We present a computational motion analysis algorithm that calculates (1) the number of layers (2) the motion of each layer and (3) the assignment of locations to layers. We show that a simple model based on few assumptions can account for a number of seemingly unrelated percepts – from transparency in plaids to motion capture in smooth contours, without the need for stimulus specific heuristics. We discuss the shortcomings of this simple model and illustrate how it can be extended to incorporate static constraints on grouping.

3.1 Introduction

In order to reliably estimate the motion of a surface, the visual system needs to combine multiple measurements across space. Often, the motion can not be reliably estimated using only the local information. This difficulty arises from the ambiguity of individual velocity measurements which may give only a partial constraint on the unknown motion (Wallach, 1935) , i.e. the "aperture problem", (Horn and Schunck, 1981; Adelson and Movshon, 1982; Marr and Ullman, 1981). To solve this problem, most models assume a two stage scheme whereby local readings are first computed, and then integrated in a second stage to produce velocity estimates. Psychophysical (Adelson and Movshon, 1982; Movshon et al., 1986; Welch, 1989) and neurophysiological (Movshon et al., 1986; Rodman and Albright, 1989) findings are consistent with such a model.

Hildreth (1983) and others (Horn and Schunck, 1981; Grzywacz and Yuille, 1991) have suggested models whereby the aperture problem is solved by assuming smoothness of the velocity field, i.e. by assuming that adjacent points have similar velocities. In such models the predicted velocity field is found by minimizing a cost functional that has two terms — a "data" term that enforces that the velocity fields satisfy the local motion constraints, and a "smoothness" term that penalizes velocity fields that have large derivatives. Poggio et al. (85) have shown that the smoothness assumption is useful in many aspects of computational vision. They pointed out that many problems in vision are "ill-posed" in the mathematical sense — there are not enough constraints in the data to arrive at a reliable solution. Solving ill-posed problems by adding an additional smoothness constraint is common in applied mathematics and is known as "regularization" (Tikhonov and Arsenin, 1977). Marroquin et al. (1987) described a Bayesian interpretation of the functional minimized in regularization, where the smoothness term corresponds to a prior favoring smooth velocities. Recently, we have shown (Weiss and Adelson, 1998) that a Bayesian motion estimator with a prior favoring slow and smooth velocity fields is consistent with a large number of published phenomena in human motion perception when the scene contains a single motion.

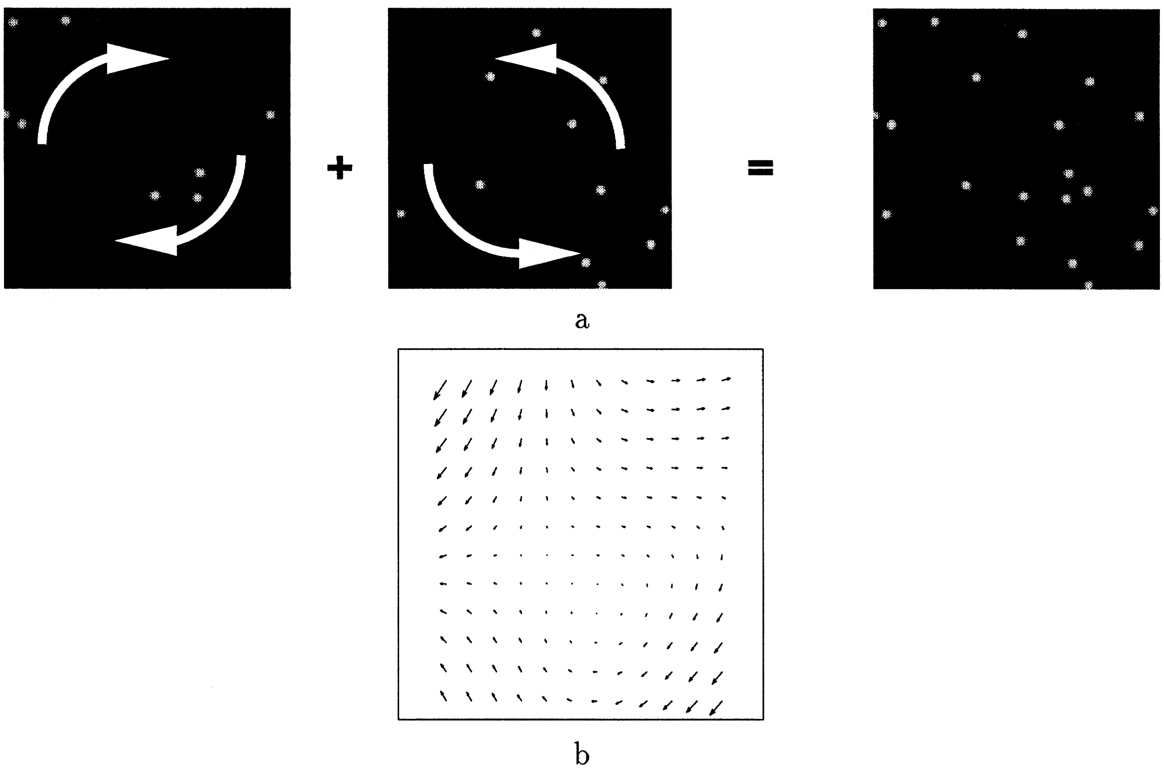


Figure 3-1: **a.** Two transparent rotating sheets rotating in opposite directions form a single sequence. Humans perceiving this scene see the two rotations. **b.** The output of a standard smoothness algorithm on this sequence. The algorithm tries to simultaneously fit the motion of both surfaces and recovers a single elastic deformation that is not at all like the human percept.

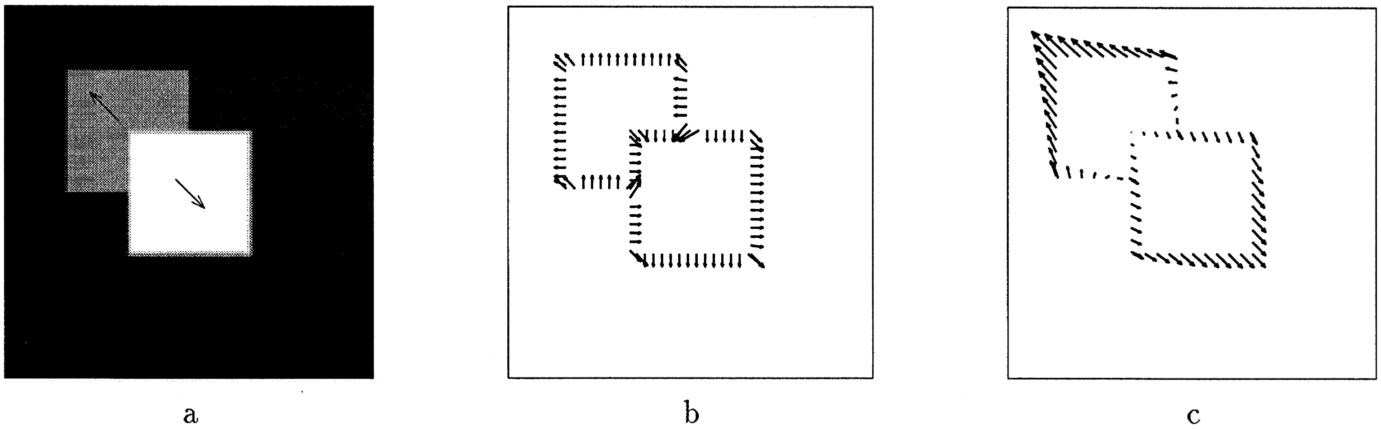


Figure 3-2: **a.** Two squares translate in the image in different directions. **b.** The output of a local motion analyzer on this scene. The estimate is correct at the corners of the squares but two types of errors can be seen in the estimated motion. First, at the straight edges of the squares the aperture problem is encountered, there is not enough local information to determine the correct motion and normal velocity is predicted. Second, at the junctions formed where one square occludes the other, there is unambiguous motion information that is accidental — the motion of these junctions is not related to the motion of either of the two squares. **c.** The output of a standard smoothness algorithm on this sequence. The algorithm tries to simultaneously fit the motion of both surfaces and recovers a single elastic deformation that is not at all like the human percept.

In scenes containing multiple motions, however, the smoothness assumption by itself gives estimates that is nothing like human perception. Figure 3-1 shows an example with two transparent surfaces rotating in opposite directions. Figure 3-1b shows the output of a standard smoothness based algorithm (Grzywacz and Yuille, 1991) on this stimulus. The estimated velocity field tries to simultaneously fit both motions. While humans looking at such a scene perceive two transparent surfaces each with its own smooth motion field, the algorithm predicts a single, irregular velocity field.

A second example of the problems associated with global smoothness is shown in figure 3-2a. Two squares translate rigidly in opposite directions. Figure 3-2b shows the output of a local motion analyzer on this scene (Lucas and Kanade, 1981; Bulthoff et al., 1989). The estimate is correct at the corners of the squares but two types of errors can be seen in the estimated motion. First, at the straight edges of the squares the aperture problem is encountered, there is not enough local information to determine the correct motion and normal velocity is predicted. Second, at the junctions formed where one square occludes the other, there is unambiguous motion information that is accidental — the motion of these junctions is not related to the motion of either of the two squares. Figure 3-2c shows the output of a global smoothness algorithm (Grzywacz and Yuille, 1991). Since it pools all measurements together, the algorithm predicts a single elastic deformation rather than two rigid translations.

The failure of global smoothness algorithms in scenes containing multiple motions such as figure 3-1 is well known and several ways of fixing the smoothness assumption have been proposed. Hildreth (1983) proposed a model whereby smoothness is only assumed along contours. Her algorithm found the velocity field of least variation along the zero crossings of the image. To illustrate her assumption consider the two squares scene discussed in figure 3-2. Hildreth's algorithm would first extract contours from this scene and then combine measurements along the contour. Thus assuming that the first step correctly extracted two contours, one for the boundary of each square, her algorithm would only assume smoothness in the motion of each square. It would not assume any relationship between the motions of the two squares. Thus for this stimulus it would predict two rigid motions, consistent with human perception.

Although Hildreth's assumption of smoothness along contours does solve some of the problems associated with smoothness models, there is reason to believe it is not exactly the assumption used by the human visual system. As pointed out by Gryzywacz and Yuille (1991) the Hildreth assumption would predict no influence between features that are off the contour and the perceived motion of the contour. This is inconsistent with experimental results that show a strong influence of features in such displays (e.g.(Nakayama and Silverman, 1988a; Shiffrar et al., 1995; Weiss and Adelson, 1995; Rubin and Hochstein, 1993)). Figure 3-3 shows an example dating back to Wallach (1935). A line whose endpoints are invisible appears to move in the normal direction, but when a small number of dots translating horizontally are added to the display they tend to "capture" the line, and the line appears to move horizontally. The fact that the dot influences the line when it is not part of the line's contour is inconsistent with Hildreth's model or any other model that assumes smoothness only along contours.

An alternative approach to "fixing" the problems associated with global smoothness assumption was to assume piecewise smoothness, or smoothness with discontinuities. In these models, e.g. (Terzopoulos, 1986; Hutchinson et al., 1988; Horn, 1986), nearby points are assumed to have similar velocities, but if the velocities are too dissimilar the assumption is abandoned and a discontinuity is assumed there instead. An advantage of these models over the standard smoothness models is that when the location of the discontinuity is estimated correctly, there is no smoothing across boundaries. This avoids many of the oversmoothing problems associated with global smoothness algorithms.

Despite these successes, the piecewise smoothness assumption can not directly account for the percept of transparency(cf. (Marroquin, 1992; Darrell and Pentland, 1991; Madrasmi et al., 1993)). For example, in the scene with two transparent rotating sheets shown in figure 3-1, such algorithms would estimate a single velocity field with many discontinuities. This does not seem to capture the salient part of the human

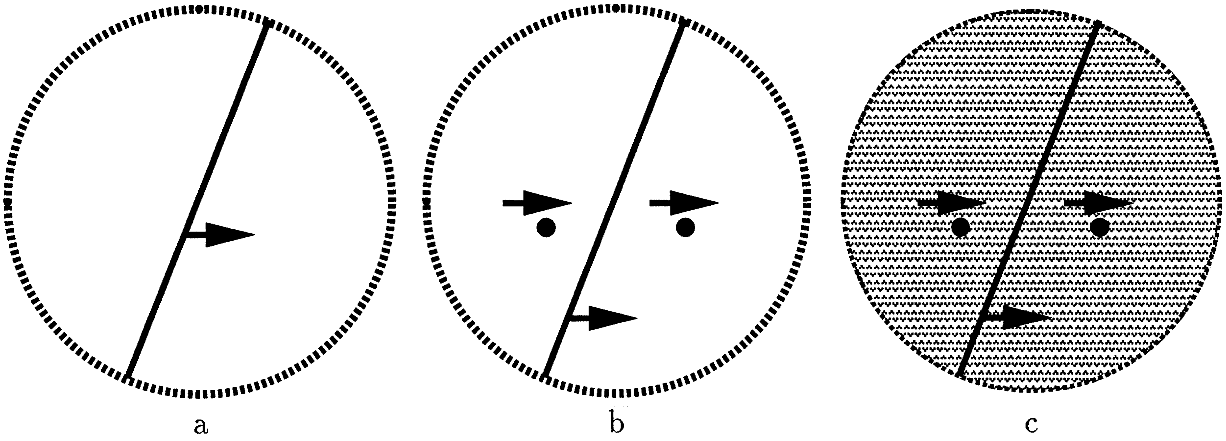


Figure 3-3: **a.** A horizontally translating diagonal line whose endpoints are invisible is consistent with an infinite family of motions. Typically, under these conditions, the normal velocity is chosen and the line appears to translate diagonally. (Wallach 35) **b.** When two horizontally translating dots are added to the display the line appears to move in the direction of the dots (Wallach 35, Rubin and Hochstein 93). This is inconsistent with a model that only combines information along contours (e.g. Hildreth 83). **c.** The effect persists when the display is placed on a static texture background. This is inconsistent with an algorithm that assumes "smoothness with discontinuities" (e.g. Terzopoulos 86). The discontinuities formed between the dots and the background would inhibit any interactions between the dots and the line.

percept in such scenes — humans tend to report seeing two surfaces, each moving with a smooth motion. The fact that there are two global surfaces is simply not part of the vocabulary used by the piecewise smooth models.

In addition to the failure to account for the percept of surfaces, the discontinuities approach also predicts an incorrect motion for certain scenes. Essentially, it predicts no interaction between two locations if there is a motion discontinuity between them. Figure 3-3c shows a simple example in which the line and the dot translate horizontally over a static background.The dissimilarity between the motions of the dots and the background texture would give rise to a discontinuity as would the dissimilarity between the line and the texture. Yet human perceiving this scene report no difference between the percept with and without the static texture. The dots and the line appear to be in front of the texture and are perceived as a single surface. Thus while piecewise smoothness may be a reasonable assumption to make in many contexts, it does not appear to be sufficient for modeling human motion perception.

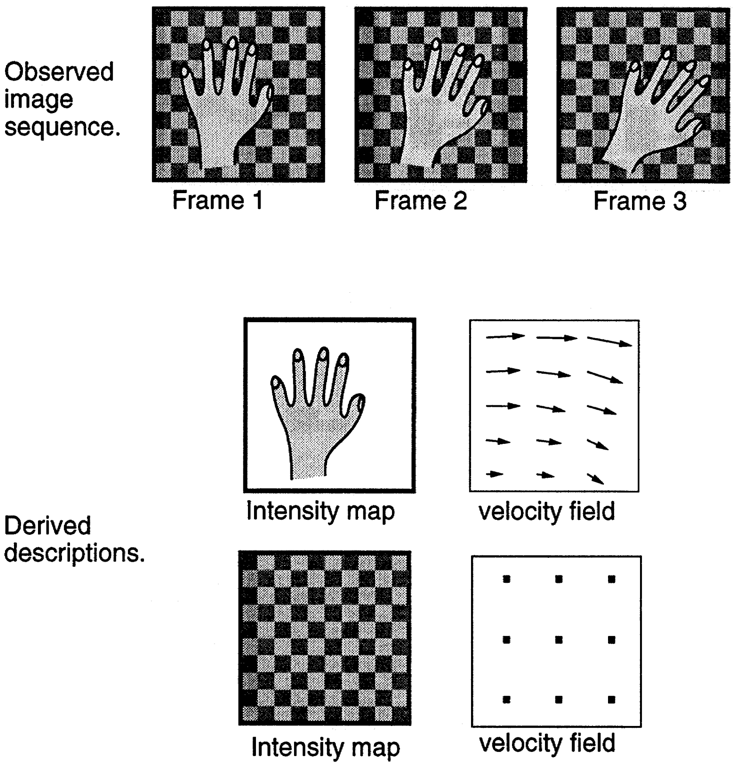


Figure 3-4: Layered decomposition of image sequences (adapted from (Wang and Adelson, 1994)). In a layered description, an image sequence is decomposed into a small number of occluding layers or surfaces, and each layer has a corresponding motion field. In this paper we propose that human motion perception assumes the motion field of each layer is smooth, but does not assume smoothness between motion fields of different layers.

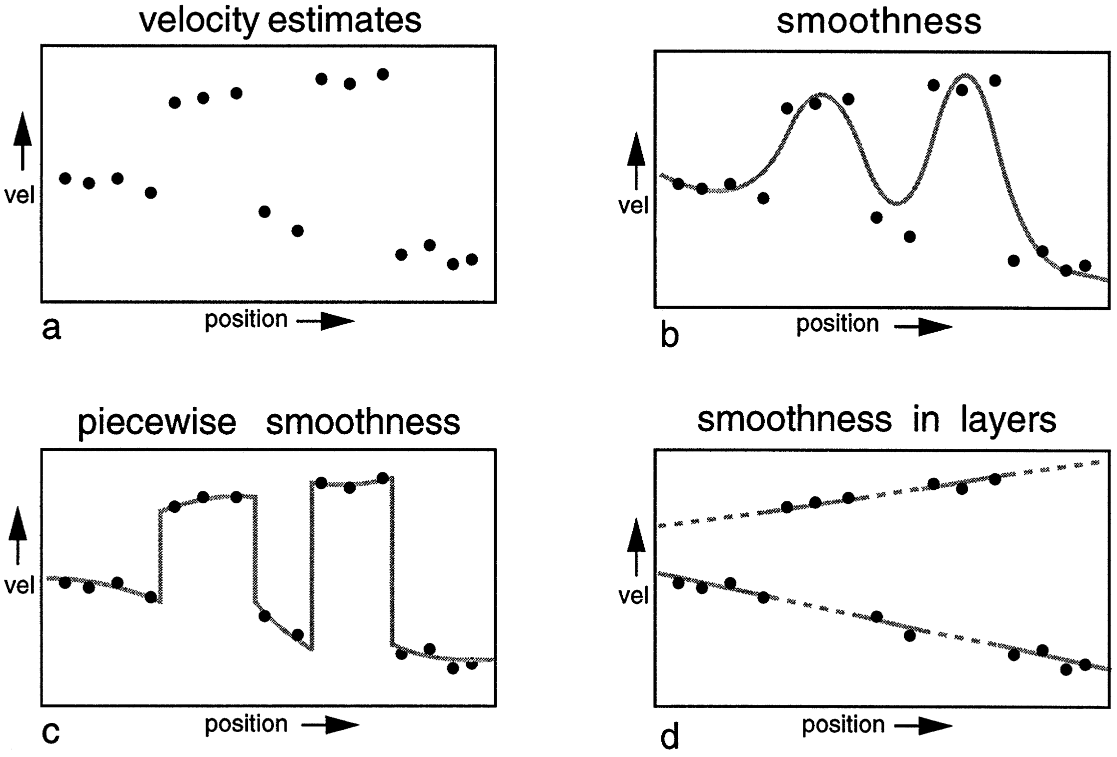


Figure 3-5: An illustration of the smoothness in layers assumption in $1D$ (adapted from (Wang and Adelson, 1994)). **a.** Hypothetical velocity estimates as a function of position. Such data would typically arise from two surfaces in depth. **b.** Global smoothness assumption applied to this data. The measurements from the two surfaces are mixed together rather than segmented. **c.** Piecewise smoothness. Information is not propagated across discontinuities. The resulting estimate is rather noisy. **d.** Smoothness in Layers. Two smooth velocity functions are found, one for each surface.

As these simple demonstrations show, the visual system does not appear to assume global smoothness over the image, nor does it assume smoothness only along contours, nor does it assume smoothness with discontinuities. In this paper we propose a formulation that we call "smoothness in layers". We assume the scene includes a small number of surfaces or layers (Wang and Adelson, 1994) and that motion varies smoothly within a given layer. To illustrate this assumption consider figure 3-4. Global smoothness would assume that motion varies smoothly over the entire image, while smoothness in layers assumes that one velocity field will vary smoothly over the front surface and a second velocity field will vary smoothly over the back surface. There is *no* assumption of smoothness between two layers only within layers.

Unfortunately, the input to the visual system is not a description in terms of surfaces or layers. Thus if we wish to account for human motion perception by assuming smoothness in layers, we need to also account for the formation of a layered description from spatiotemporal data. In this paper we present a computational model that does precisely that.

The model presented here is a computational theory in the sense of Marr and Poggio (Marr, 1982; Marr and Poggio, 1977) who emphasized the different levels of understanding at which perception can be investigated. Rather than describing a particular biological implementation of a particular algorithm, we attempt to find the constraints and assumptions used by the visual system when estimating motion. Specifically, we employ a Bayesian framework and search for a probability model such that the most probable interpretation under that model will correspond to the percept seen by humans. In order to test such a model we need a way of (1) specifying the probability of an interpretation given a gray level image sequence and (2) finding the most probable interpretation. The validity of the model is obtained by comparing the most probable interpretation to human perception of the same sequence. It is in this sense that the models of Hildreth (1983) or Gryzywacz and Yuille (1991) are not satisfactory — in scenes containing multiple motions the most probable interpretation given their assumptions do not correspond to the human percept.

The subsequent section describes mixture estimation, the general statistical framework that we will to formulate our assumptions. Using that framework we then formulate a simple model that is based on a minimum number of assumptions and see the extent to which such a model can explain a range of percepts. The model receives as input a gray level image sequence and calculates (1) the number of layers (2) the assignment of pixels to layers and (3) the velocity field of each layer. It is built on the following assumptions: (1) a preference for a small number of layers and (2) a preference for slow and smooth velocity fields within a layer. Surprisingly, we find that these two assumptions can account for a large number of percepts ranging from the tendency of plaid patterns to cohere to the rigid and nonrigid percepts of smooth contours and dots. We then show, however, that this model is not sufficient to account for other percepts, particularly due to a failure to combine form and motion constraints. We end by discussing how these additional constraints can be incorporated into the same statistical framework.

3.2 Mixture models for motion analysis

3.2.1 An illustration of mixture estimation

Mixture estimation refers to the problem of inferring the parameters of multiple processes from data. This framework served as the basis for the "mixture-of-experts" architecture in neural networks (e.g. (Jordan and Jacobs, 1994; Jacobs et al., 1991)) and has become increasingly popular in computer vision for segmentation problems (e.g. (Jepson and Black, 1993; Ayer and Sawhney, 1995)). To understand the ideas behind mixture estimation we illustrate it first using an abstract problem — the fitting of lines to data. Figure 3-6 illustrates the difference between mixture estimation and the more commonly known method of regression. In regression, one is given noisy data that was generated by a single line and the problem is to find the parameters of the best fitting line. In mixture estimation, one is given noisy data generated by multiple lines and the problem is to find (1) the parameters of the best fitting lines and (2) the assignment of datapoints to a given line.

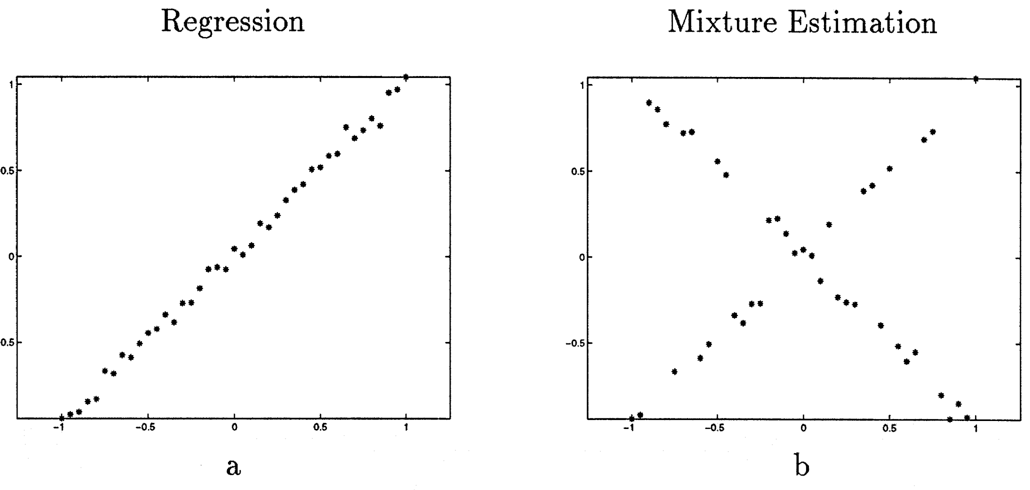


Figure 3-6: To convert the "smoothness in layers" notion into a computational model, we use the framework of mixture models. Here we illustrate mixture models in the familiar setting of line fitting. In regression we are given data generated by a single line with added noise, and we are trying to estimate the parameters of the line. In mixture estimation, we are given data generated by two lines with additive noise and we are trying to find (1) the assignment of points to lines and (2) the parameters of the two lines. This second problem is analogous to motion segmentation.

Formally, in regression problems the assumption is that the points (x_i, y_i) satisfy the equation:

$$y_i = ax_i + b + \sigma\nu \tag{3.1}$$

where ν is normally distributed Gaussian random noise.

In a two component mixture model the assumption is that the points satisfy:

$$y_i = \begin{array}{ll} a_1x_i + b_1 + \sigma\nu & , L_i = (1,0) \\ a_2x_i + b_2 + \sigma\nu & , L_i = (0,1) \end{array} \tag{3.2}$$

where L_i is an indicator variable that determines whether point (x_i, y_i) was generated by line 1 or line 2.

The well known least squares equations for regression arise from assuming the generative model in equation 3.1 and finding the most likely parameter values for $\Theta = (a, b)$ given the data. Thus in regression one maximizes:

$$P(\{x_i, y_i\}|\Theta) = \alpha \prod_i e^{-(ax_i+b-y_i)^2/2\sigma^2} \tag{3.3}$$

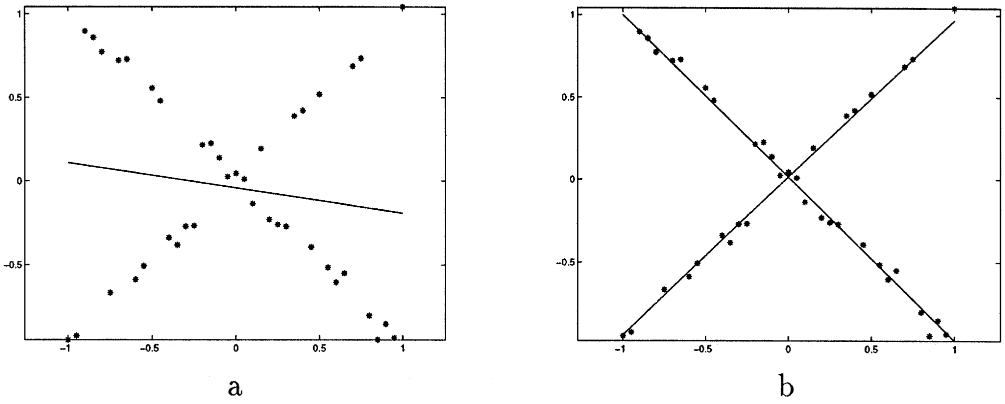


Figure 3-7: **a.** The result of performing regression on the data generated by a mixture model. **b.** The result of performing mixture estimation on the same data. Similar to the situation in motion estimation, good estimates for the parameters of the line requires combining the information from many points while avoiding the mixing together of points that belong to different lines.

(throughout this paper we adopt the convention of writing probability distributions with a normalizing constant α that guarantees that the distribution integrate to 1).

Similarly, one can find the most likely parameter values $\Theta = [(a_1, b_1), (a_2, b_2)]$ given the data assuming the generative model in equation 3.2. To maximize the likelihood we need to maximize:

$$P(\{x_i, y_i\}|\Theta) = \alpha \prod_i \sum_{j=1}^{2} e^{-(a_j x_i + b_j - y_i)^2 / 2\sigma^2} \tag{3.4}$$

Figure 3-7a shows the best fitting single line for the data (obtained by maximizing equation 3.3) while figure 3-7b shows the best fitting pair of lines (obtained by maximizing equation 3.4.

Unlike regression, the mixture likelihood function cannot be maximized analytically. Typically, an iterative algorithm known as the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is used. This algorithm is based on the observation that mixture estimation can be thought of as the solution of two subproblems: (1) the estimation of the hidden labels L_i and (2) the estimation of the parameters of the two lines. The intuition behind EM is that each subproblem is easy to solve assuming the other one is solved. That is, assuming we know the assignment of each datapoint,

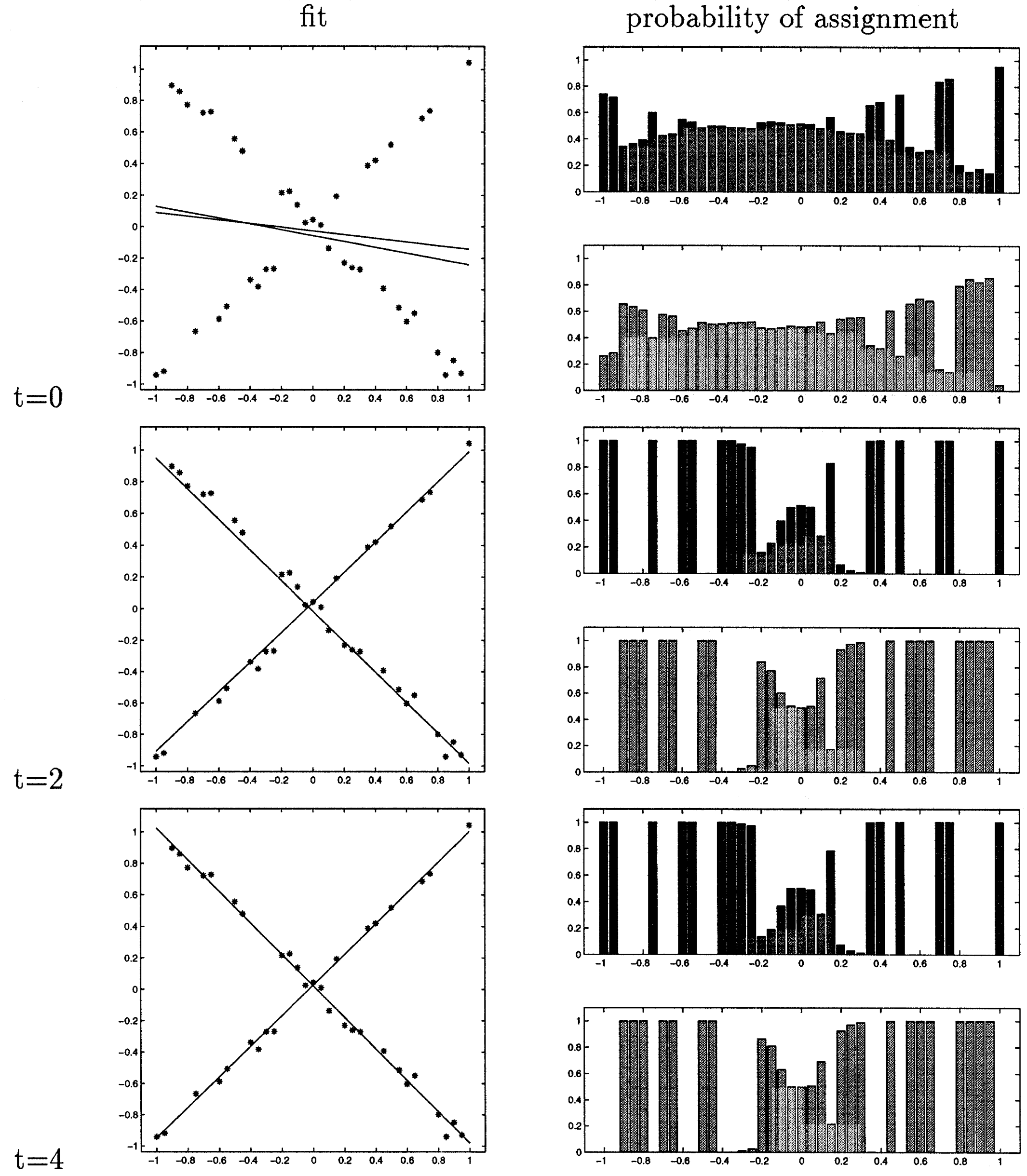


Figure 3-8: An illustration of the Expectation-Maximization (EM) algorithm on the data in figure 3-6. The algorithm iteratively updates the parameters of the two lines (middle column) and the probability of assignment of points to lines (rightmost column). Each panel in the rightmost panel contains two bar graphs — the probability of assignment of points (as a function of x location) to line 1 and 2 respectively. Note that for a fixed x location, the probabilities sum to 1. The iterations are guaranteed to increase the likelihood at every iteration.

then we can estimate the parameters of each line by taking into consideration only those points assigned to it. Likewise, if we know the parameters of the lines we can assign each point to the line that fits it best.

This gives the basic structure of an EM algorithm:

- start with random parameter values for the two models.
- Iterate until parameter values converge:
  - E step: assign points to the model that fits it best.
  - M step: update the parameters of the models using only points assigned to it.

In fact both steps are slightly more complicated, due to the assignment being continuous rather than binary valued. In the E step the probability of every point being assigned to a particular line is estimated, and in the *M* step these probabilities are used as weights in weighted regression for the parameters of the models. The algorithm is guaranteed to increase the likelihood at every iteration, and will typically converge to a local maximum of the likelihood. The update rules for the EM algorithm with line fitting are given in the appendix. Figure 3-8 shows an example run. The algorithm is initialized with random initial conditions and iteratively estimates the probability of assignment (right column) and the line fitting (middle column). Each panel in the rightmost panel contains two bar graphs — the probability of assignment of points (as a function of x location) to line 1 and 2 respectively. Note that for a fixed x location, the probabilities sum to 1.

Figure 3-7 illustrates some of the commonalities between motion analysis and line fitting in the presence of noise. Due to the noise in the data, a reliable estimate for the parameters of the line requires using as many data points as possible. At the same time, using all the data points will give the wrong answer. This is analogous to the problem the visual system faces when estimating motion in scenes containing multiple motions — due to the aperture problem a reliable estimate for motion requires using as many locations as possible, but using all the locations (as in global smoothness) will give the wrong answer. This analogy suggests that mixture estimation may be a useful framework for modeling human motion analysis. Before describing the smoothness in layers motion analysis model, we first describe two variants on mixture estimation — the use of prior probabilities on the curves and the estimation of the number of models.

3.2.2 Mixture of smooth curves

In the previous example we used mixture estimation to fit multiple straight lines to data. The framework, however, can also be used to fit multiple curves to data. Thus, rather than using straight lines one could use parabolas, or higher order polynomials. These approaches would restrict the possible curves that one can fit to a small parametric family. Alternatively, one could fit multiple smooth curves to the data, where smoothness of a curve is defined in terms of a derivative operator.

An elegant result that can be derived in the regularization framework (e.g. (Girosi et al., 1995)) shows that for many smoothness definitions, the optimal curve can be expressed as a superposition of basis functions. A commonly used smoothness operator yields the Gaussian basis function expansion:

$$y(x;\theta) = \sum_j \theta_j G(x - x_j) \tag{3.5}$$

where $G(x)$ is a Gaussian and θ_j describe the particular curve. The mixture model now becomes:

$$y_i = \begin{array}{ll} \sum_j \theta_j^1 G(x_i - x_j) + \sigma\nu & , L_i = (1,0) \\ \sum_j \theta_j^2 G(x_i - x_j) + \sigma\nu & , L_i = (0,1) \end{array} \tag{3.6}$$

where L_i is an indicator variable that determines whether point (x_i, y_i) was generated by curve 1 or curve 2. The likelihood function is now:

$$P(\{x_i, y_i\}|\Theta) = \alpha \prod_i \sum_{k=1}^{2} e^{-(\sum_j \theta_j^k G(x_i - x_j) - y_i)^2 / 2\sigma^2} \tag{3.7}$$

Note that the number of parameters describing a single curve $\{\theta_j\}$ is equal to the number of datapoints. Thus there are twice as many unknowns as there are datapoints in the mixture model. In order to make the estimation well defined, we introduce a *prior probability* over curves $y(x)$:

$$P(y(x)) = \alpha e^{-J(y)} \tag{3.8}$$

with:

$$J(y) = \int \|Dy(x)\|^2 dx \tag{3.9}$$

where D is a differential operator (e.g. the first derivative, second derivative etc.). We can use this to define a prior probability over the parameters:

$$P(\theta) = P(y(x;\theta)) \tag{3.10}$$

Now rather than maximizing the likelihood of the parameters Θ, we can maximize their posterior probability:

$$P(\Theta|\{x_i, y_i\}) = \alpha P(\{x_i, y_i\}|\Theta)P(\theta^1)P(\theta^2) \tag{3.11}$$

Figure 3-9 shows the result of maximizing equation 3.11 for the data in figure 3-6. Even though the data can, in principle, be fit with a single, oscillating curve, the two relatively straight curves have higher posterior probability.

3.2.3 Estimating the number of components

The mixture likelihood equations (equations 3.4– 3.7) assume two components in the mixture. How can this number be estimated from the data?

Consider the data in figures 3-6a-b. The data in figure 3-6a was generated by a single line, while that in figure 3-6b was generated by a mixture of two lines. Suppose we are trying to decide for each dataset whether to fit the data with a single line or with two lines. A possible approach is to compare the likelihoods of the two solutions. A first problem arises from the fact that equation 3.4 is by definition a function of

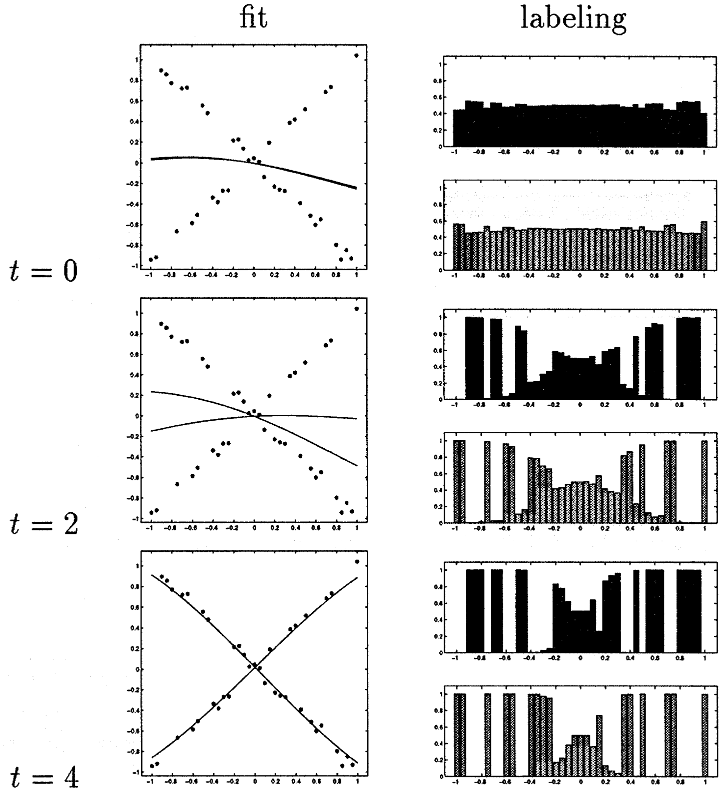


Figure 3-9: The results of running the EM algorithm with arbitrary smooth curves on the data in figure 3-6. Each curve has as many degrees of freedom as there are datapoints and hence the entire data set can in principle be fit by a single, oscillating curve. However, the fit with two curves maximizes the posterior probability

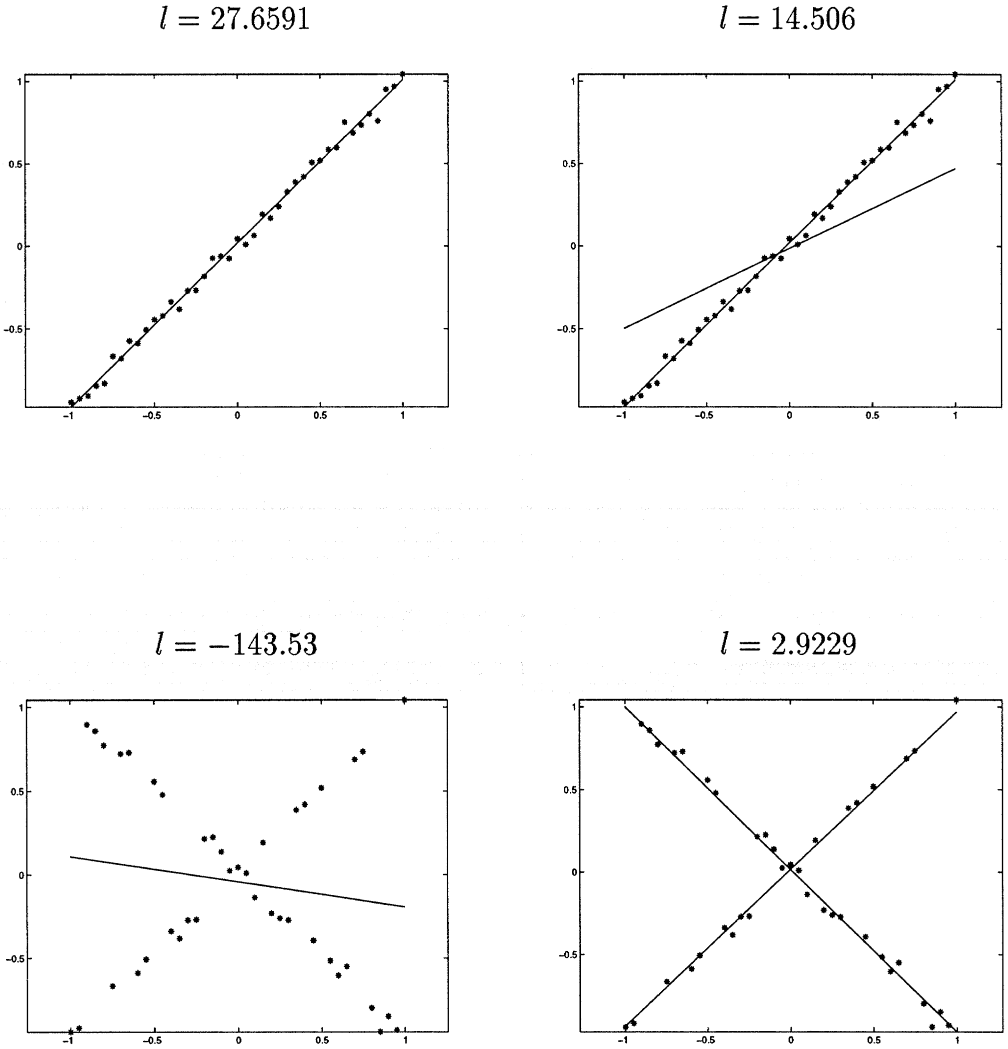
four parameters (corresponding to two numbers for each line). This would suggest that we need to calculate a separate likelihood for the one line case, using a mixture model with a single component. It turns out, however, that the likelihood for a single line in a one component mixture is equal by definition to the likelihood of two lines in a two component mixture where both lines have the same parameters[1]. Thus to calculate the likelihood of a single line with parameters (a, b) we can substitute $(a_1 = a, b_1 = b), (a_2 = a, b_2 = b)$ into equation 3.4.

A second problem is that it may seem that we need to add an additional "complexity" term to equation 3.4 in order to solve this problem. Given that equation 3.4 depends on the residual between the fit of the lines and the data, it seems that one can always obtain a better fit using two lines rather than one.

In the absence of noise $\sigma \rightarrow 0$ this intuition is correct — the maximum likelihood solution will always contain two distinct lines rather than one. However for nonzero σ the intuition is misleading. The likelihood may be maximized with a single distinct line. In (Weiss, 1998) we derive analytical results regarding the conditions in which one line maximizes the likelihood rather than multiple lines (see also (Rose et al., 1990; Durbin et al., 1989; Tenenbaum and Todorov, 1995) for a similar analysis in other contexts). Roughly speaking, if the fit with a single likelihood has mean squared residual less than σ then a single line will be preferred over multiple lines. The intuition behind this is that additional lines will be fitting the noise and not the data. Figure 3-10 illustrates this idea. It shows the relative likelihoods of one or two lines for the two datasets in figure 3-6. The parameter σ is held constant throughout. Note that for data generated by a single one line, a solution with two lines actually has lower likelihood. For the data generated by two lines, the solution with one line has lower likelihood. Thus for a nonzero σ the likelihood function can be used to compare solutions with a different number of lines.

To summarize: mixture estimation provides a way to fit multiple models to data in a principled statistical framework. Given data generated by multiple processes and a probabilistic generative model it enables the calculation of (1) the number of

[1] the normalizing factor α absorbs the difference.



$l = 27.6591$ $\quad$ $l = 14.506$

$l = -143.53$ $\quad$ $l = 2.9229$

Figure 3-10: In mixture models with known σ the number of lines can be estimated by maximizing the likelihood. Here we show the log likelihoods (equation 3.4 with $\alpha = 1$) for two datasets and various solutions. For data generated by a single line (top panels) the fit with a single line is favored over the one that uses two lines. For data generated by two lines (bottom panels) the likelihood of a two line fit is higher than the one that only uses a single line. For these examples, there is no need to add an additional "complexity cost" to the likelihood. Analytical results showing this can be found in (Weiss, 1998).

processes (2) the probability of a particular datapoint being generated by each process and (3) the parameters of the processes. In the next section we describe how this framework can be used to formalize the idea of smoothness in layers.

3.2.4 The smoothness in layers mixture model

In line fitting, the problem was to account for the datapoints $\{x_i, y_i\}$ with a small number of lines. In our model we perform motion analysis in a similar way – the problem is to account for the local motion data with a small number of surface motions.

What do we mean by local motion data? One possibility would be to extract a velocity vector locally and try to account for that velocity. However, such an approach would ignore the fact that local motion data has varying degrees of ambiguity — from ambiguous edges in which only one component of the local velocity can be estimated to corners in which both components can be estimated. In (Weiss and Adelson, 1998) we used the following likelihood model for the velocity at a given point:

$$L(v_x, v_y) = P(I_x, I_y, I_t | v_x, v_y) = \alpha e^{-C(v_x, v_y)/2\sigma_N^2} \qquad (3.12)$$

where $C(x, y)$ quantifies the degree of consistency of the velocity with the local data. It is based on the gradient constraint (Horn and Schunck, 1981; Lucas and Kanade, 1981):

$$C(v_x, v_y) = \sum_{x,y,t} w(x, y, t)(I_x v_x + I_y v_y + I_t)^2 \qquad (3.13)$$

where v_x, v_y denote the horizontal and vertical components of the local velocity I_x, I_y, I_t denote the spatial and temporal derivatives of the intensity function and $w(x, y, t)$ is a spatiotemporal window centered at (x, y, t). The gradient constraint is closely related to more physiologically plausible methods for motion analysis such as autocorrelation and motion energy (Reichardt, 1961; Poggio and Reichardt, 1973; Adelson and Bergen, 1986; Simoncelli, 1993). As we have shown in (Weiss and Adelson, 1998) this local likelihood can describe a large range of local motion measurements with different amounts of ambiguity.

Equation 3.12 is the analogue of the regression likelihood (equation 3.3). It gives the likelihood for a single velocity at location (x, y). When we are dealing with a mixture model, say with two velocities $(v_x^1, v_y^1), (v_x^2, v_y^2)$ the likelihood becomes:

$$L(v_x^1, v_y^1) = \alpha \sum_{j=1}^{2} e^{-C(v_x^j, v_y^j)/2\sigma_N^2} \tag{3.14}$$

This is the analogue of the mixture likelihood (equation 3.4) in the line fitting case.

As in (Weiss and Adelson, 1998) we use a 50 dimensional space to represent the motion of a layer. The mapping from parameter space to the velocity field is given by:

$$v_x^j(x, y) = \sum_{i=1}^{25} \theta_i^j G(x - x_i, y - y_i) \tag{3.15}$$

$$v_y^j(x, y) = \sum_{i=26}^{50} \theta_i^j G(x - x_i, y - y_i) \tag{3.16}$$

where $G(x, y)$ is a two dimensional Gaussian function in image space, with spatial extent defined by σ_x:

$$G(x, y) = e^{-\frac{x^2+y^2}{2\sigma_x^2}} \tag{3.17}$$

and (x_i, y_i) define a $5x5$ grid that covers the image. Note that this parameterization is the analogue of equation 3.5 in the curve fitting case.

We also include a prior probability over the parameters θ. Following (Weiss and Adelson, 1998) we use a prior favoring slow and smooth velocity fields. Formally, we define the following prior on a velocity field, $V(x, y)$:

$$P(V) = \alpha e^{-J(V)/2\sigma_P^2} \tag{3.18}$$

with:

$$J(V) = \sum_{xy} \|Dv(x, y)\|^2 \tag{3.19}$$

here Dv is a differential operator, i.e. it measures the derivatives of the velocity field. We follow Grzywacz and Yuille (1991) in using a differential operator that penalizes velocity fields with strong derivatives:

$$Dv = \sum_{n=0}^{\infty} a_n \frac{\partial^n}{\partial x} v \tag{3.20}$$

These equations are the analogue of equations 3.8–3.10 in the smooth curve fitting case.

Thus we assume that the spatiotemporal data was generated according to the following model. First, for each layer, a velocity field $v^j(x, y)$ is drawn from the distribution favoring slow and smooth velocity fields. Then a labeling of the image is generated, i.e. a vector $L(x, y)$ at every location such that $L_k(x, y) = 1$ if and only if position x, y will be assigned to group k. Given the labelings and the velocity field, the likelihood of the observation at location (x, y) is given by equation 3.12.

To perform inference in this model, we need to find the values of θ that maximizes the posterior probability. Again, this can be done with the Expectation-Maximization algorithm. In the E step, we calculate $\hat{L}_k(x, y)$ — the probability that the derivatives at location (x, y) are due to layer motion k. In the M step we use these probabilities as weights in a weighted least squares problem for θ^k. Since EM may converge to a local maximum, we use multiple restarts and choose the one that has higher posterior probability. The update rules are given in the appendix.

The generative model has two important constants. σ_N and σ_R which describe the standard deviations of the data noise and the prior respectively. The ratio between these two numbers will determine the tradeoff between accounting for the data and the smoothness assumption. If σ_N is much smaller than σ_R then the MAP estimate will fit the data well but may not be smooth. The absolute size of σ_N will influence the number of models in the MAP estimate. As mentioned in the previous sections, for $\sigma_N \rightarrow 0$ the posterior will typically be maximized with a large number of distinct layers. However for nonzero σ_N the posterior may be maximized with a smaller number of layers that avoid overfitting. Given these two constants and the spatiotemporal intensity function we can calculate (1) the number of layers (2) the motion of each layer and (3) the probability of a measurement being assigned to a layer.

3.3 Results of the simple model

The generative model described in the previous section attempts to account for the motion data using smooth velocity fields. It can be thought of as having two main assumptions (1) preference for a small number of layers (embodied in the nonzero σ_N) and (2) smooth and slow velocity fields within a layer (embodied in the finite σ_P). In this section we examine the extent to which the most probable interpretation of a scene when using these assumptions correspond to human percepts. We refer to this model as the simple smoothness in layers model, or simple SIL model.

To understand the output of the model, we first show the output on the stimuli discussed in the introduction — two transparent sheets in rotation, two rigidly translating squares, and a diagonal line translating horizontally with added dots. The algorithm used only two frames to calculate the spatio-temporal derivatives and the window used to calculate the motion measurements was of size $3x3$ pixels. The images were $64x64$ pixels. For computational efficiency, the algorithm was restricted to find at most two layers, but it had to decide between one or two layers based on the posterior probability. To avoid edge artifacts, all measurements obtained at the edges of the images were discarded.

Figure 3-11 shows the output of the simple SIL model on the two rotations scene. The output consists of (1) the number of layers (2) the motion of each layer and (3) the probability of a measurement being assigned to a given layer. The right hand column shows the two velocity fields estimated by the model. Unlike the global smoothness output shown in the introduction, the two velocity fields have a strong rotational component. The left hand column shows the probabilities of assignment displayed as a gray level image. White pixels correspond to high probability, black correspond to low probability and gray pixels denote intermediate probabilities. Note that all locations that do not have any motion information (i.e $3x3$ windows that have no spatial gradients) the probability is gray. That is because such regions, based purely on motion constraints, are equally likely to belong to either layer. The measurements around the dots, however, are classified as belonging to one rotation or the other.

Figure 3-12 shows the output on the two squares scene. The algorithm finds the correct two translations and estimates the probabilities of all measurements belonging to each of the translations. Note the grayness of all pixels in the interiors of the squares — since there is no motion information there, there is equal probability they belong to either of the two layers. The difficulty of the problem is perhaps more evident when considering the output of the local motion analyzer shown in figure 3-2b. Note that most of the local motions are either pure horizontal or pure vertical motion and yet the algorithm is not confused by this fact. This is because it does not attempt to simply fit the local motion data as is, but rather takes into account the local uncertainty. Thus the algorithm knows that the many horizontal and vertical motions signals have a high degree of local ambiguity — unlike the motion measurements derived at the corners. Note also that the algorithm does not choose the two translations of the junctions, but rather the correct two translations. This is simply a result of the fact that there are more corners than accidental junctions — the posterior probability of the interpretation with the correct two translations is higher. Note also that the two velocity fields are highly smooth translations, rather than the relatively nonsmooth deformation estimated with the global smoothness algorithm. Since the algorithm is based on the assumption of smoothness within a layer, the two layered description is favored over the single layer description.

Figure 3-13 shows the output on the diagonal line sequence. The model segments the line from the static background and the motion of the line is estimated as diagonal. This is because of the prior probability favoring slow and smooth velocity fields within a layer. For a layer that only includes a single line, the slowest velocity field is the normal velocity. This result is identical to that reported in (Weiss and Adelson, 1998) where a single motion field was assumed. Here, however, the line is automatically segmented from the background. Figure 3-14 shows the output when two translating horizontal dots are added to the display. The algorithm again finds two layers — one corresponding to the static texture and the other corresponding to the dots and the line. Note that now, the moving layer is assumed to be moving horizontally. Thus the motion of the line is now predicted to be horizontal rather than diagonal. This

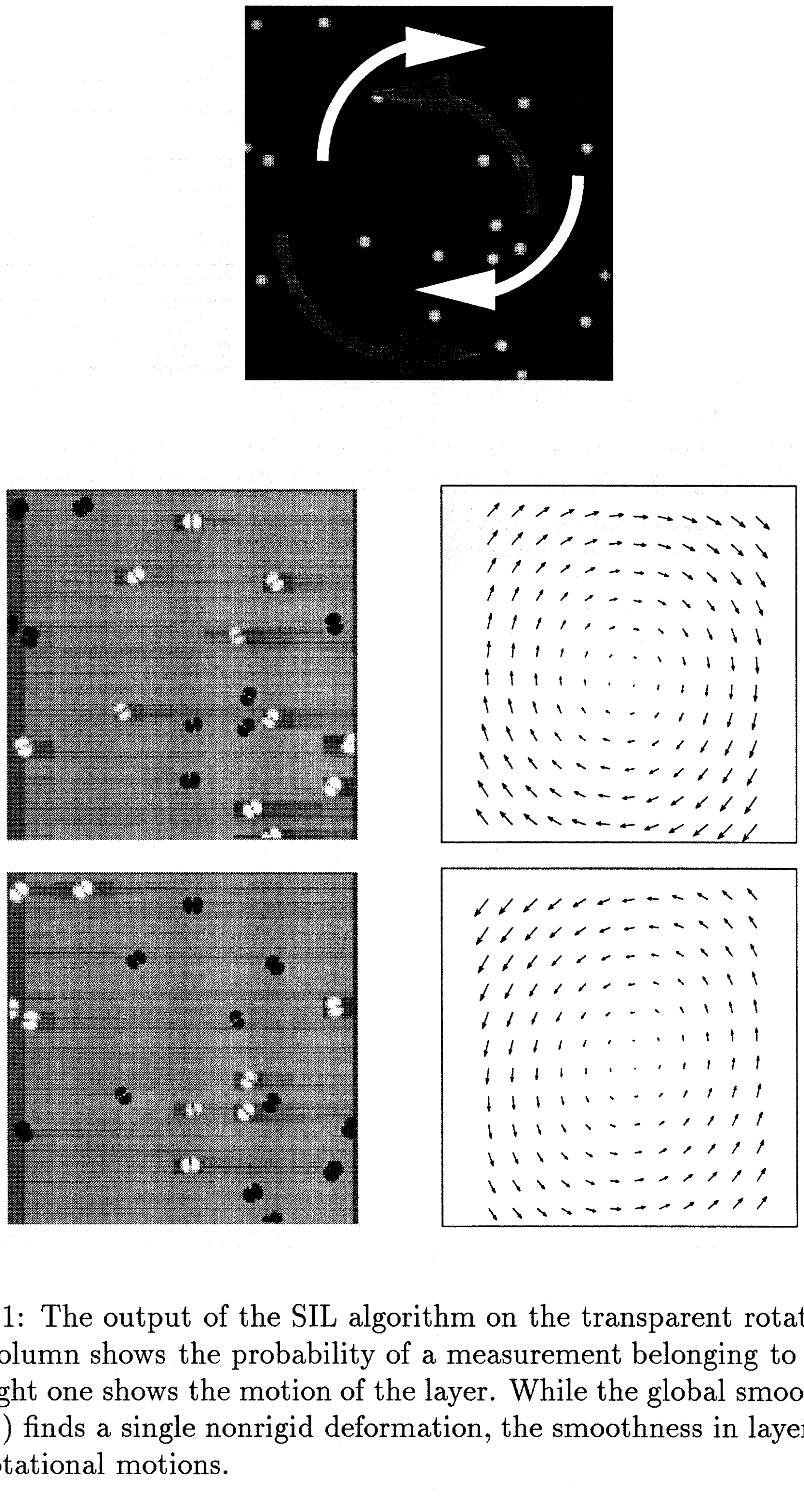


Figure 3-11: The output of the SIL algorithm on the transparent rotation sequence. The left column shows the probability of a measurement belonging to a given layer, and the right one shows the motion of the layer. While the global smoothness output (figure 3-1) finds a single nonrigid deformation, the smoothness in layers model finds the two rotational motions.

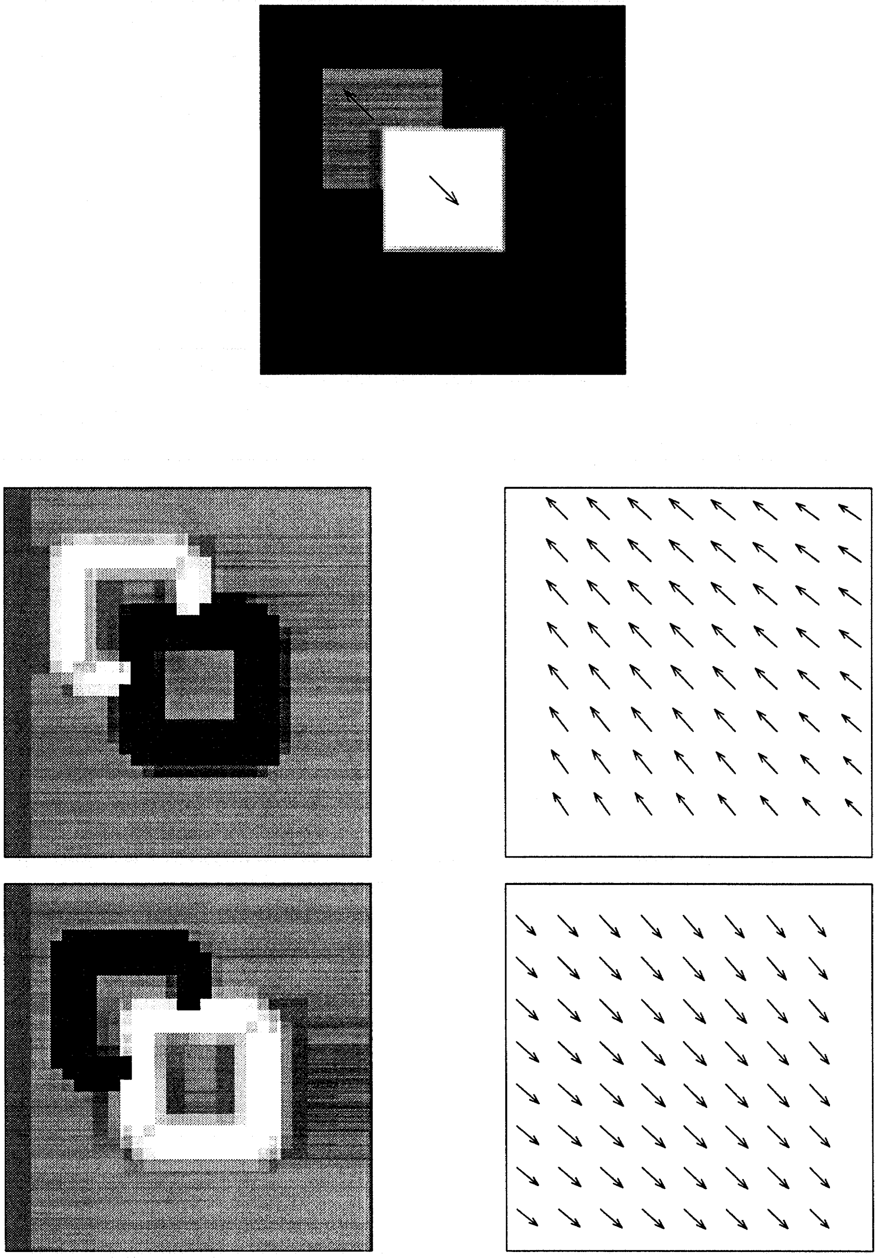


Figure 3-12: The output of the SIL algorithm on the two squares sequence. The left column shows the probability of a measurement belonging to a given layer, and the right one shows the motion of the layer. While the global smoothness algorithm (figure 3-2c) finds a single elastic deformation, the SIL finds two smooth motion fields

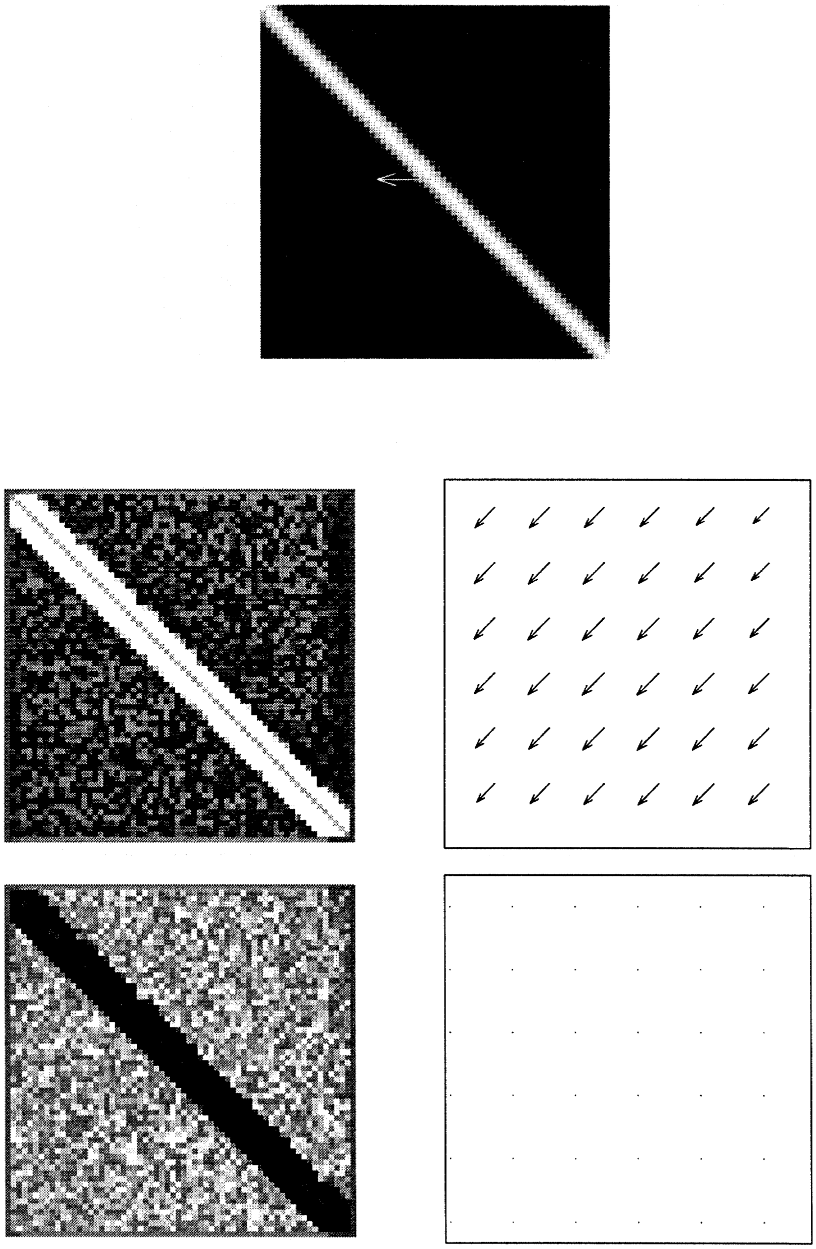


Figure 3-13: The output of the SIL algorithm on the plain translating line case. The left column shows the probability of a measurement belonging to a given layer, and the right one shows the motion of the layer. The line is segmented from the static texture background and the line is predicted to move in the normal direction. This is due to the prior favoring slow and smooth motions within a layer

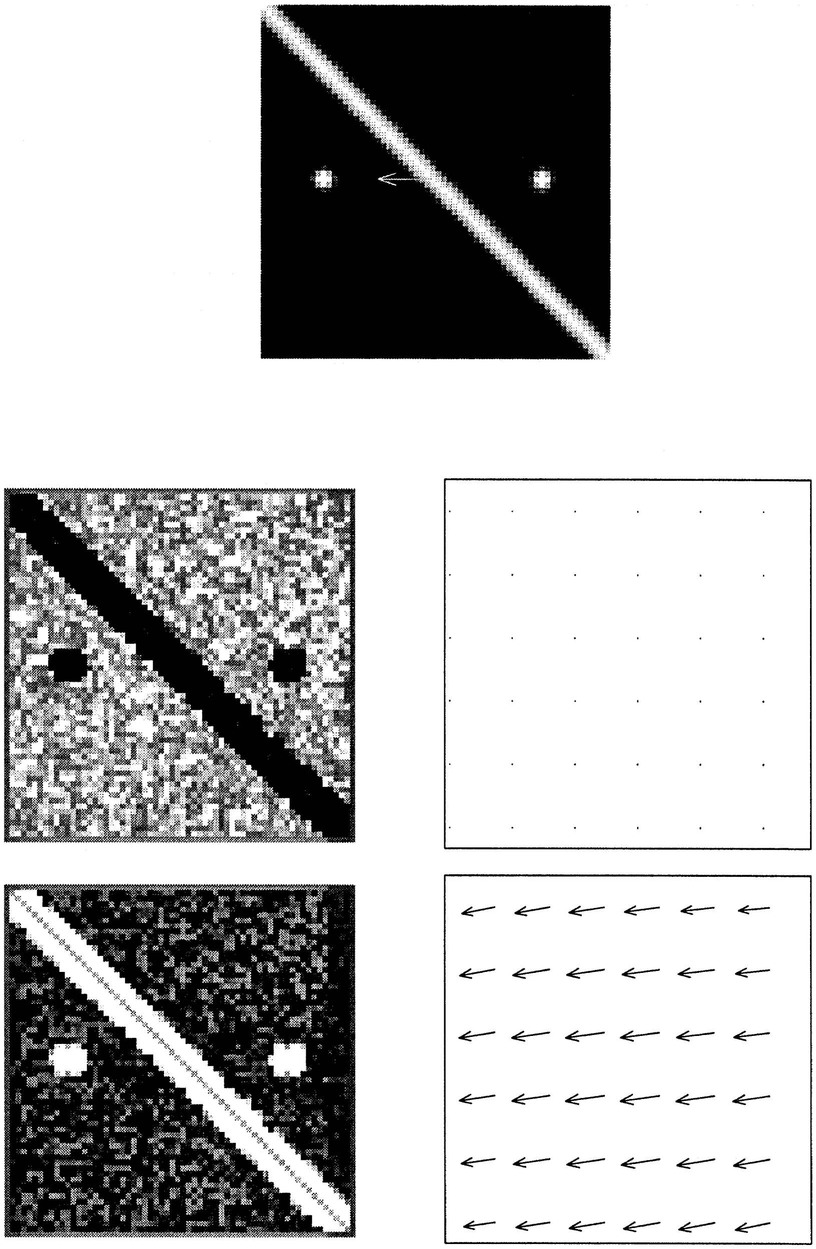


Figure 3-14: The output of the SIL algorithm on the line and dot scene. The left column shows the probability of a measurement belonging to a given layer, and the right one shows the motion of the layer. The line is segmented from the static texture background and grouped together with the moving dots. Thus the line is predicted to move in the horizontal direction, consistent with human perception.

type of influence of features off the contour on the motion of the contour is a basic prediction of the SIL approach in contrast to the smoothness along contours approach of Hildreth (1983) that does not predict such an influence. Note also that the presence of static texture between the dot and the line does not disable this influence. This is in contrast to the piecewise smoothness algorithms (e.g. (Terzopoulos, 1986)) that would predict no influence of the dot motion on the line motion in this case.

These simple stimuli show the advantage of the SIL framework – unlike global smoothness algorithms it can handle scenes containing multiple motions, but unlike the smoothness along contours or piecewise smoothness it can account for the influence of features off the contour on the motion of a contour. Note, however that these results depend somewhat on the parameter settings σ_P, σ_N that determine the tradeoff between preference towards slow and smooth on one hand, and preference towards a small number of models. Obviously, there exist parameter settings for which the algorithm will always find a single motion and reduce to the global smoothness model. On the other hand, there exist parameter settings for which the scene will be segmented into many layers, each of which moves very slowly and smoothly. In that case, the model will never give capture. The SIL framework would be more convincing if stimulus manipulations that affect the tendency to see one or two motions in humans would have a similar effect on the model when the parameters are held constant. In the next section, we examine some previously examined phenomena and compare them to the SIL model.

3.3.1 The tendency towards coherence in plaids

Probably the most well studied stimulus in which the question of one motion versus two has been studied is the "plaid" stimulus composed of two oriented gratings in motion. Since the days of Musatti (1924) (Musatti, 1924; Wallach, 1935) it was known that such patterns can be perceived in two very different ways. Either both gratings are perceived as moving coherently in a single direction, or they can be seen as sliding over each other. Adelson and Movshon (1982) investigated systematically the conditions under which one percept is preferred over another. Using sine wave gratings that were combined in an additive fashion, they found that the two gratings needed to have different contrasts for the percept of transparency to occur. The amount of contrast difference needed "decreased as the speed of the components grating increased, as the angle between their primary directions increased and as the difference between their spatial frequencies increased." The influence of speed and primary directions was also observed in square wave plaids (Farid and Simoncelli, 1994), in addition to a number of static transparency cues. Stoner, Albright and Ramachandran (1990) manipulated the luminance of the "diamond" junctions formed when the two bars intersect. When the luminances were consistent with static transparency subjects were more likely to see two component motions rather than one coherent motion. Bresssan et al. (1993) manipulated the widths and the luminances of the two square waves and found that component motion was more common when the static cues were consistent with occlusion.

Thus the tendency of plaids to cohere provides a rich set of phenomena to test the SIL model. Although we can not expect the simple model to be influenced by subtle form manipulations that influence the static segmentation, we wanted to see how much of the motion related phenomena can be captured with the simple SIL model.

3.3.2 Effect of component direction on coherence of plaids

Phenomena: Adelson and Movshon (1982) found that the tendency of two gratings to cohere decreased as the angle between their primary direction increased. This result was replicated in (Kim and Wilson, 1993) and extended to square wave plaids in (Farid and Simoncelli, 1994). To illustrate this tendency consider the two plaid stimuli in figure 3-15. The principal directions of the gratings are the direction at which each grating would appear to move when displayed by itself and moving with the pattern velocity. Thus for the plaid in figure 3-15a these two directions are $+70, -70$ while for the one in figure 3-15b the directions are $+20, -20$. Hence the plaid in figure 3-15b is more likely to be perceived as coherent. Obviously this tendency has nothing to do with static cues for transparency — if the same two plaids translate vertically rather

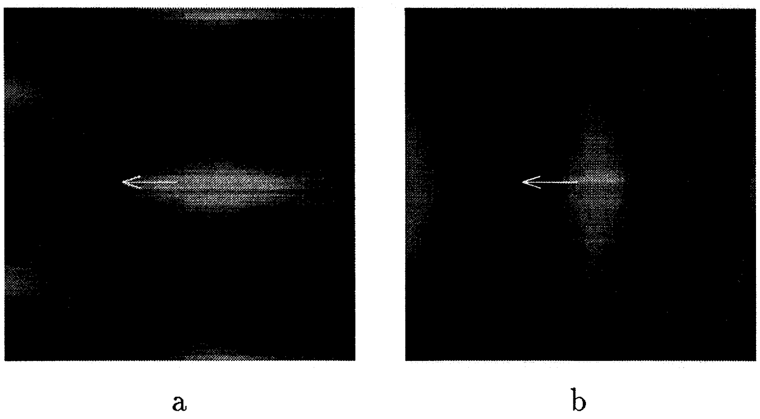


Figure 3-15: Adelson and Movshon (1982) found that the tendency of plaids to cohere depended on the difference between the principal direction of the two gratings. Thus the plaid in **a** tends to cohere less than the plaid in **b.**

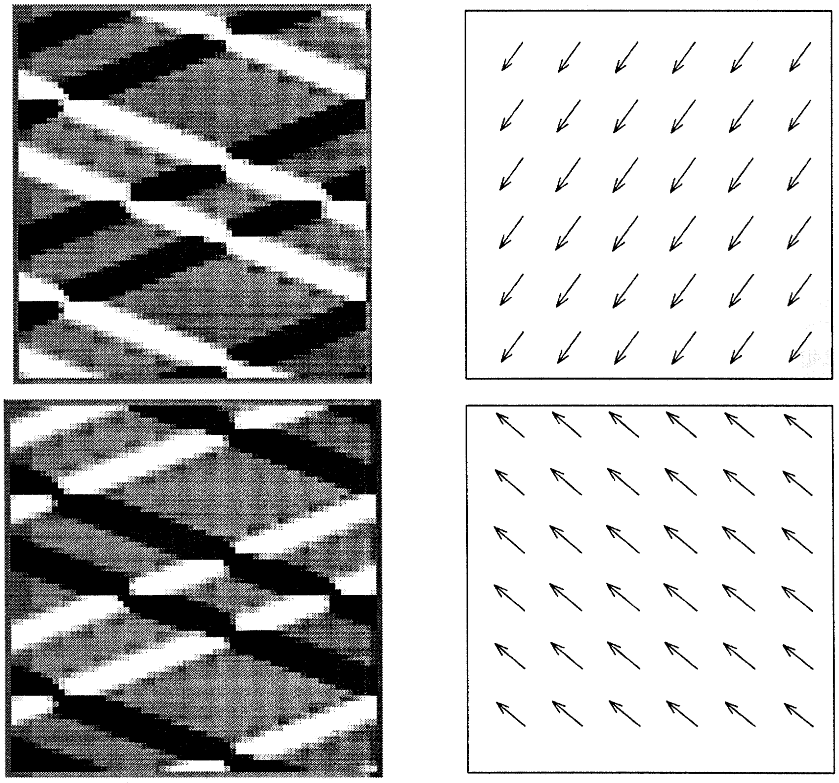


Figure 3-16: The output of the SIL algorithm on the sequence in 3-15a. The algorithm finds two layers, one corresponding to each grating, and each moves with the approximate normal motion.

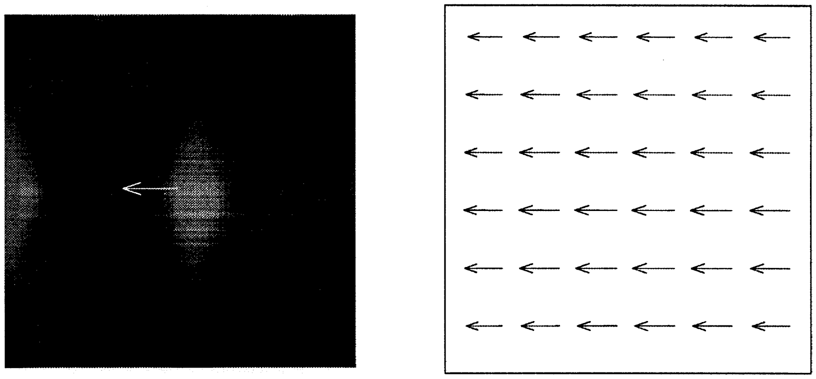


Figure 3-17: The output of the SIL algorithm on the sequence in 3-15b. The algorithm finds a single layer corresponding to the whole plaid and whose motion is the pattern motion. The parameters of the algorithm are identical to those used in figure 3-16. Thus the algorithm's tendency to see coherence in plaids depends on the differences between the principal directions of the gratings, similar to human observers.

than horizontally the tendency to cohere is reversed and the one in figure 3-15a is more likely to be perceived as coherent.

Model Results: Figure 3-16 shows the output of the SIL model on the stimulus in figure 3-15a. The algorithm finds two layers — one corresponding to each grating. The motion of each layer is roughly the normal velocity of the respective gratings. Figure 3-17 shows the output of the SIL model with identical parameters on the stimulus in figure 3-15a. Now a single layer is found corresponding to the plaid and its motion is horizontal. An additional layer (not shown) was found with static motion. This layer was assigned only two pixels in the entire image, pixels that due to aliasing artifacts were not consistent with the plaid motion.

Discussion: This result can be interpreted as a tradeoff between the preference for a small number of layers and the preference towards slow and smooth motions. The plaid motion is always faster than the component motions, hence if the system only had a preference towards slow speeds the plaid would never cohere. On the other hand, the coherent motion uses only a single layer so if the system only had a preference towards a small number of layers, plaids would always cohere. When the angle between the gratings is varied, the difference between the plaid speed and the component speed varies as well. Assuming the plaid speed is held constant at

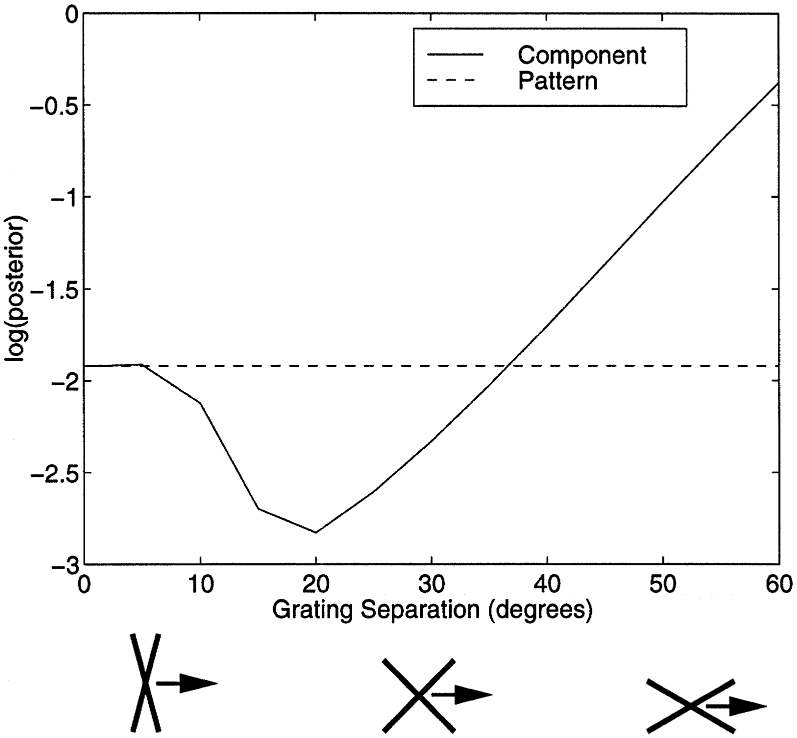


Figure 3-18: The posterior probabilities of the coherent and transparent percepts in the case of an ideal plaid stimulus as a function of the angle between the two components. For large difference between the two principal directions, the normal velocities are much slower than the pattern velocities and hence the transparent percept has higher posterior probability. For small (nonzero) differences between the component motions, the difference in speeds is negligible and the preference for a small number of layers causes the coherent percept to be preferred. When the two angles are identical, the coherent and transparent interpretation-s are identical.

1, the component speed is $cos(\alpha/2)$ where α is the angle between the two principal directions. Thus for the plaid in figure 3-15a the pattern speed is much faster than the component speeds, while for the plaid in figure 3-15b the the pattern speed is only slightly faster. Farid et al (Farid et al., 1995) also pointed out that the effect of angle on coherence may be a result of the bias towards slow speeds.

To formalize this intuition, we calculated the analytical posterior probability for an ideal square wave plaid with fixed pattern motion as the angle between the principal directions is varied. In this case the likelihood function has a particularly simple form (see appendix) and we compared the posterior probability of the coherent percept (i.e. two layers with identical pattern motion) to that of the transparent percept (two layers, each with the component motion). Figure 3-18 shows these predicted posteriors. Since the pattern velocity is held constant, the posterior probability for the coherent percept does not change when the angles are varied. The transparent percept, however, changes as a function of angle. For large difference between the two principal directions, the normal velocities are much slower than the pattern velocities and hence the transparent percept has higher posterior probability. For small differences between the component motions, the difference in speeds is negligible and the preference for a small number of layers causes the coherent percept to be preferred. When the two angles are identical, the coherent and transparent interpretations are identical. The parameters σ_N, σ_P are held constant throughout. Changing these parameters will change the crossover point of the two curves, but the qualitative behavior is unchanged.

3.3.3 Effect of pattern speed on the plaid coherence

Phenomena: For a given plaid direction and component orientation the tendency to cohere is influenced by speed. The faster the plaid is moving, the more likely it is to be perceived as transparent (Adelson and Movshon, 1982). Farid et al (Farid et al., 1995) showed that this result also holds in square wave plaids up to a certain cutoff speed — when the pattern moves too fast, subjects have difficulty in reporting their percept.

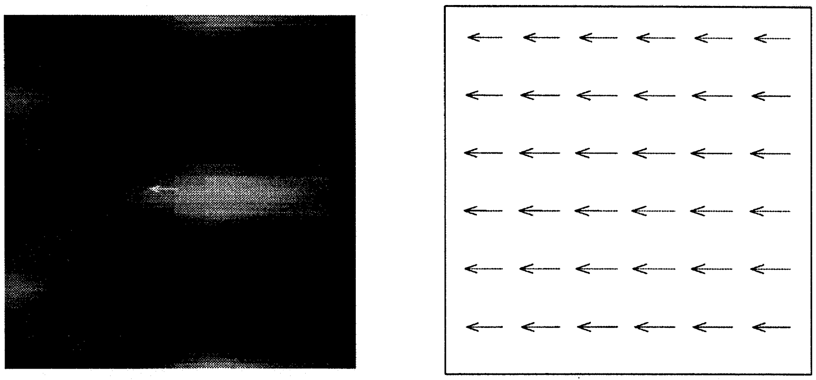


Figure 3-19: The output of the SIL algorithm on the plaid in 3-15a moving at a slower speed. Even though the parameters are held constant, the algorithm now finds a single layer corresponding to the whole plaid and whose motion is the pattern motion. Thus the algorithm's tendency to see coherence in plaids depends on the speed similar to human observers.

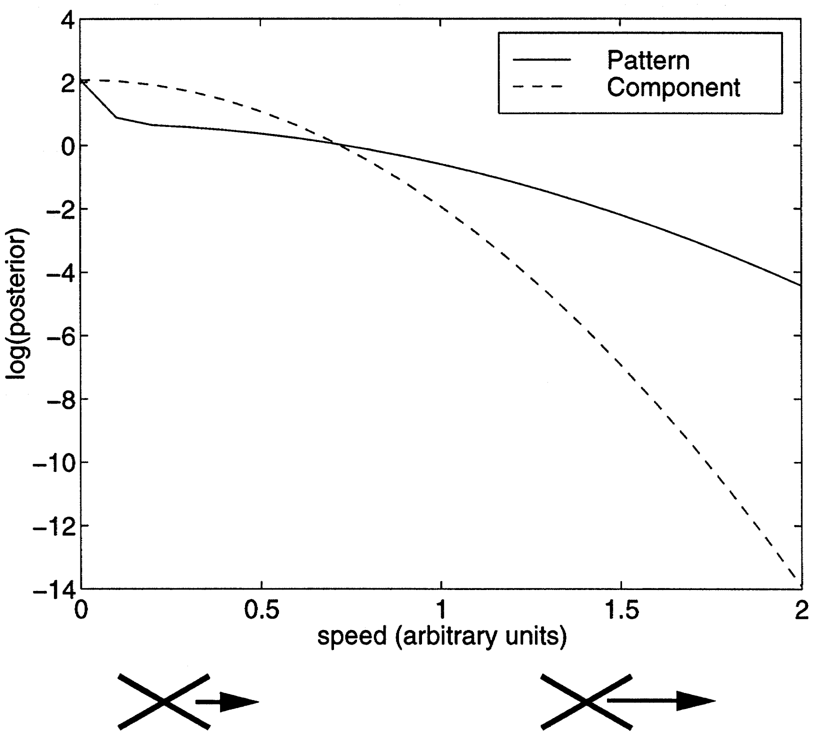


Figure 3-20: The posterior probabilities of the coherent and transparent percepts in the case of an ideal plaid stimulus as a function of the speed. For fast speeds the normal velocities are much slower than the pattern velocities and hence the transparent percept has higher posterior probability. For small (nonzero) speeds the difference in speeds is negligible and the preference for a small number of layers causes the coherent percept to be preferred. For zero speed, the two percepts are identical.

Model Results Figure 3-19 shows the output of the SIL model on the stimulus in figure 3-15a when it is moving at 1/10 the speed than in the previous section. Even though all parameters are held constant, the model now prefers a coherent percept for the plaid.

Discussion: Again this is the result of the tradeoff between number of layers and slow and smooth motions within a layer. Although the plaid speed is always faster than the component speeds the magnitude of the speed difference depends on the speed of the plaid. For slowly moving plaids, this difference is small and therefore the model prefers a single layer. However for fast plaids, the difference in speeds between the pattern speed and the component speeds causes the noncoherent percept to be preferred. Although there are two layers here, they each move much slower than the plaid does.

To formalize this intuition, we again calculated the analytical posterior probability for a plaid with fixed angles and direction as a function of speed. Figure 3-20 shows the posterior probabilities for the coherent and transparent percept. For both percepts, the posterior probability decreases with increasing speed. For fast speeds the normal velocities are much slower than the pattern velocities and hence the transparent percept has higher posterior probability. For small (nonzero) speeds the difference in speeds is negligible and the preference for a small number of layers causes the coherent percept to be preferred. For zero speed, the two percepts are identical.

3.3.4 Influence of contrast on coherence of square wave plaids

Phenomena: For square wave plaids, the tendency to cohere increases as contrast increases (Farid et al., 1995). Thus when all other parameters are held constant, the plaid in figure 3-15a will cohere more when the contrast is raised high.

Model Results: Figure 3-21 shows the output of the SIL algorithm on the plaid in 3-15a moving at a higher contrast. Even though the parameters are held constant, the algorithm now finds a single layer corresponding to the whole plaid rather than the two layers found at low contrast.

Discussion: This result can be interpreted as the tradeoff between the likelihood

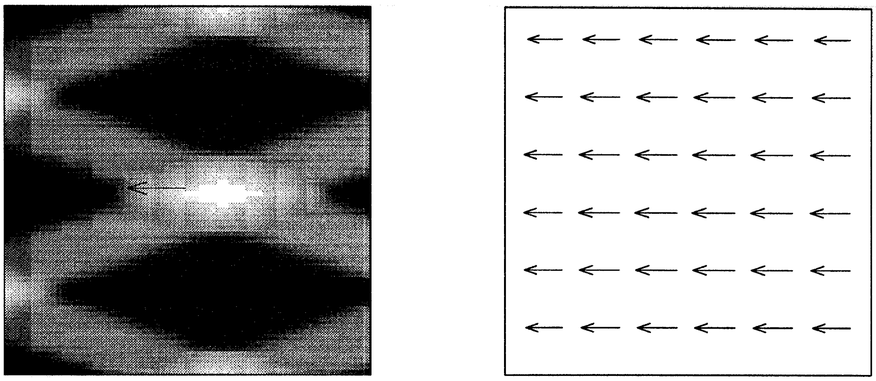


Figure 3-21: The output of the SIL algorithm on the plaid in 3-15a moving at a higher contrast. Even though the parameters are held constant, the algorithm now finds a single layer corresponding to the whole plaid and whose motion is the pattern motion. Thus the algorithm's tendency to see coherence in plaids depends on the contrast similar to human observers.

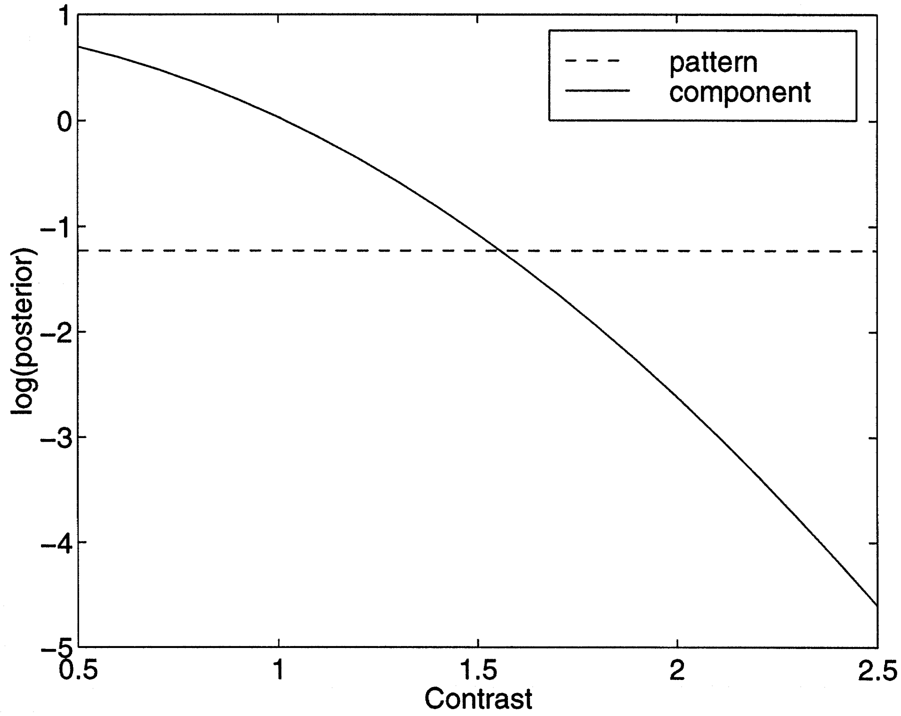


Figure 3-22: The posterior probabilities of the coherent and transparent percepts in the case of an ideal plaid stimulus as a function of the contrast. The transparent percept can account for all the measurements except those from the corners. Thus it has lower likelihood but higher prior probability (it corresponds to slower motions). At low contrast, the likelihoods are fuzzy and the prior dominates – hence the transparent percept has higher posterior probability. But at high contrasts, the likelihoods dominate and the coherent percept is preferred.

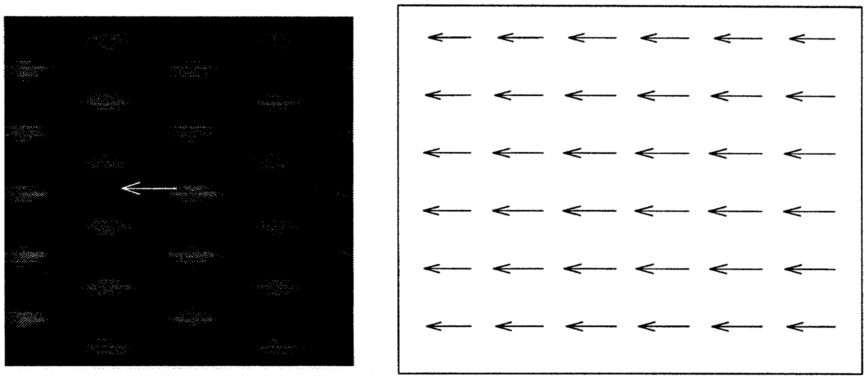


Figure 3-23: The output of the SIL algorithm on a plaid with the same orientations and speeds as that in 3-15 but at a lower period. though the parameters are held constant, the algorithm now finds a single layer corresponding to the whole plaid and whose motion is the pattern motion. Thus the algorithm's tendency to see coherence in plaids depends on the period similar to human observers.

term (explaining the data) and the posterior term favoring slow and smooth velocity fields. The transparent percept can account for all the measurements except those from the corners. Thus it has lower likelihood but higher prior probability (it corresponds to slower motions). At low contrast, the likelihoods are fuzzy and the prior dominates – hence the transparent percept has higher posterior probability. But at high contrasts, the likelihoods dominate and the coherent percept is preferred. Figure 3-22 shows the analytic posterior probability for the coherent and transparent percepts as a function of contrast. Since the coherent percept explain the data perfectly well, its posterior probability is unaffected by contrast. The transparent percept, however, does not explain the data at the corners and hence its likelihood decreases with increasing contrast. The crossover point will of course depend on the parameters σ_P, σ_N.

3.3.5 Influence of period on coherence of square wave plaids

Phenomena: For square wave plaids, the tendency to cohere decreases as period increases (Farid et al., 1995). Thus when all other parameters are held constant, the plaid in figure 3-15a will cohere more than the same plaid with higher period.

Model Results: Figure 3-23 shows the output of the SIL algorithm on a plaid

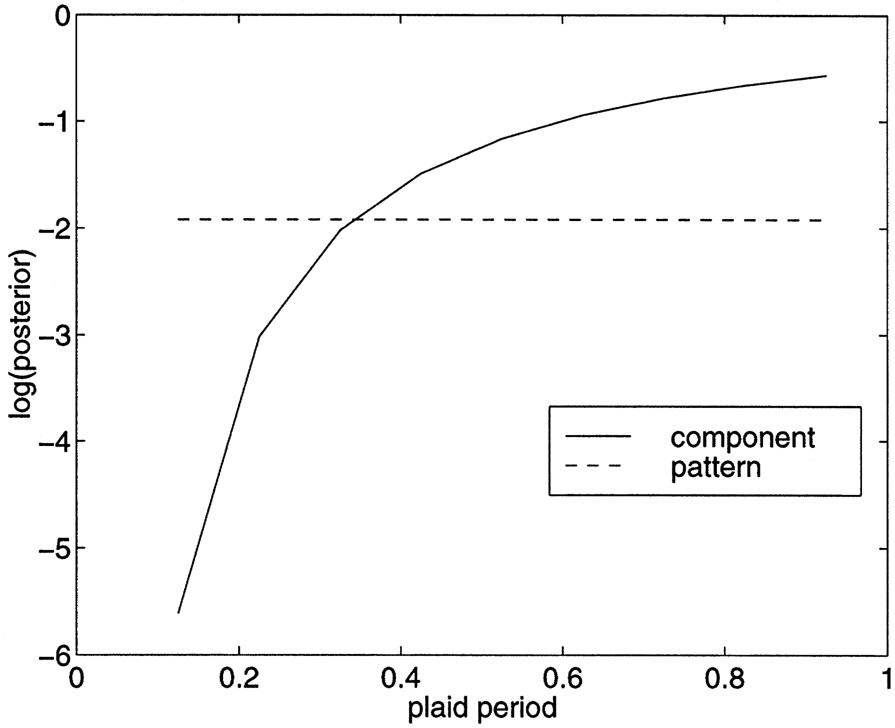


Figure 3-24: The posterior probabilities of the coherent and transparent percepts in the case of an ideal plaid stimulus as a function of the period. The transparent percept can account for all the measurements except those from the intersections. Thus it has lower likelihood but higher prior probability (it corresponds to slower motions). The number of intersections, however, changes with the period. As the period is decreased, the coherent percept is favored.

identical to that in 3-15 but with a lower period. Even though the parameters are held constant, the algorithm now finds a single layer corresponding to the whole plaid rather than the two layers found at higher period.

Discussion: This result can be interpreted as the tradeoff between the likelihood term (explaining the data) and the prior term favoring slow and smooth velocity fields. The transparent percept can account for all the measurements except those from the corners. Thus it has lower likelihood but higher prior probability (it corresponds to slower motions). The number of intersections, however, depends on the period. Thus at low periods the coherent percept is preferred. Figure 3-24 shows the analytic posterior probability for the coherent and transparent percepts as a function of period. Since the coherent percept explain the data perfectly well, its posterior probability is unaffected by the period. The transparent percept, however, does not explain the data at the corners and hence its likelihood increases with increasing period. The crossover point will of course depend on the parameters σ_P, σ_N.

3.3.6 Plaids — discussion

As the previous sections show, the tendency of the SIL model to see plaids as coherent or transparent is influenced by stimulus manipulations in a manner similar to human observers. Thus without invoking any stimulus specific heuristics, the model predicts the influence of plaid direction, speed, contrast and period on perceived coherence.

The results of the previous section were on square wave plaids. What happens when we use sine-wave plaids as stimuli? In general, the SIL model almost always sees sine-wave plaids as coherent. Since the local velocity measurements are obtained over a small local spatial patch, in sine-wave plaids all measurements have multiple orientations and are therefore only fully consistent with the pattern motion. At very low contrasts, however, and at very fast speeds, the SIL model does find two layers rather than one. At that regime, we find a dependence on principal direction and speed similar to that of square wave plaids.

There are several shortcomings of the SIL model in dealing with sine-wave plaids. First, even when two layers are found, the assignment of location to layers does not

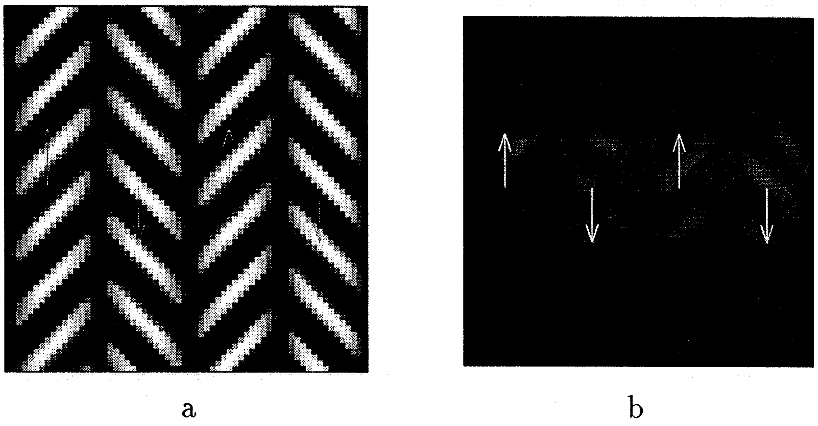


Figure 3-25: The split herringbone illusion (Adelson and Movshon, 1983). Two sets of diagonal lines translate vertically in opposite directions. At high contrast (Adelson and Movshon, 1983) the percept consists of two groups, one moving up and the other moving down. However, if the stimulus is blurred, viewed peripherally or at low contrast, one perceives a single coherent motion to the right.

reveal the two component gratings. The model attempts to assign each location to one layer or the other, but in a sine-wave plaid each location has contributions of both layers. Second, the model is not influenced by the relative spatial frequencies and contrast of the gratings in the same way that humans are. This may be similar to the situation in square wave plaids where manipulations that serve as static cues for transparency influence the human percept but are not part of the present model.

3.3.7 The split herringbone illusion

Phenomena: Adelson and Movshon (1983) conducted experiments with the "split herringbone" shown in figure 3-25. Two sets of diagonal lines translate vertically in opposite directions. At high contrast (Adelson and Movshon, 1983) the percept consists of two groups, one moving up and the other moving down. However, if the stimulus is blurred, viewed peripherally or shown at low contrast one perceives a single coherent motion to the right.

Model Results: Figure 3-26 shows the output of the SIL model on the high contrast sequence. Two layers are found and their motions are vertical. Figure 3-27 shows the output on the low contrast sequence. A single layer is found moving horizontally.

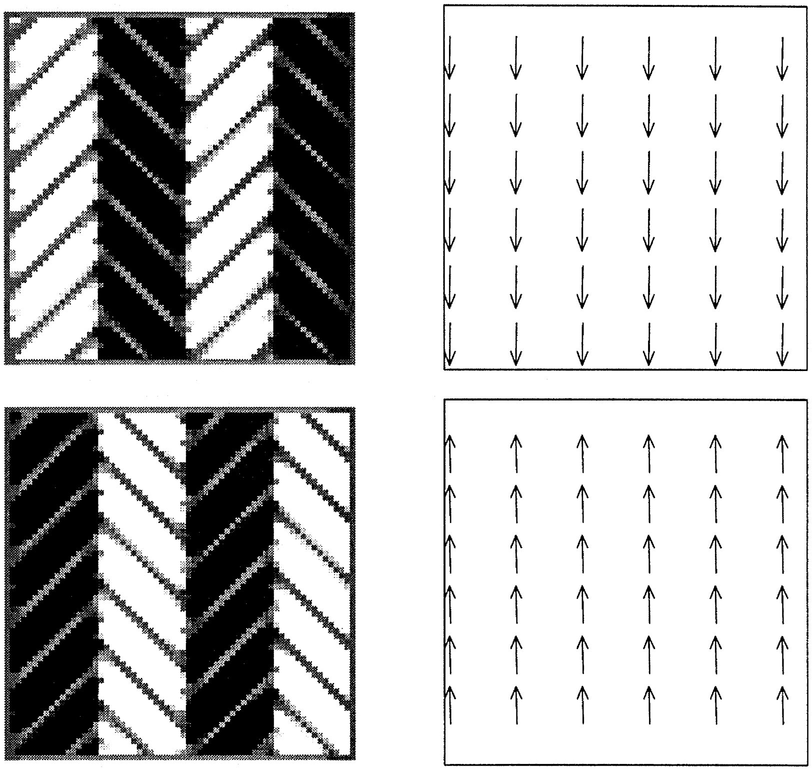


Figure 3-26: Results of the SIL algorithm on the high contrast split herringbone stimulus. Two groups are found, one corresponding to each pair of diagonal lines

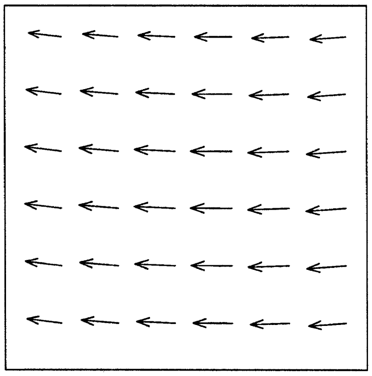


Figure 3-27: Results of the SIL algorithm on the low contrast split herringbone stimulus. One layer is found, corresponding to the full pattern. The parameters are held constant.

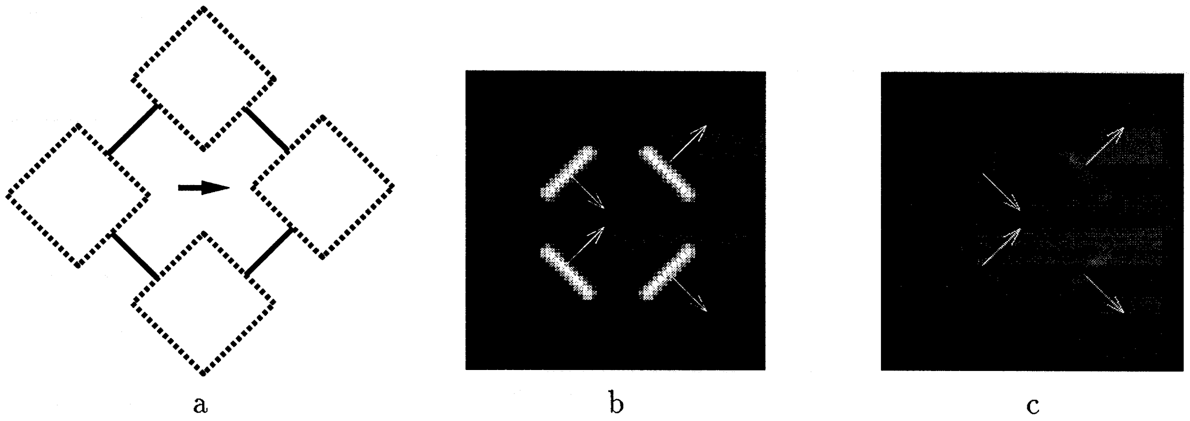


Figure 3-28: The occluded diamond (Lorenceau and Shiffrar, 1992). A diamond translates horizontally behind invisible occluders. At high contrast the percept consists of two groups, one moving up and to the right and the other moving down and to the right. However, if the stimulus is blurred or viewed peripherally or at low contrast however, one perceives a single coherent motion to the right.

Discussion: As these results show, the "illusion" of the leftward motion is actually the most probable interpretation assuming the SIL model. Recall that the model attempts to account for the motion data with a small number of slow and smooth layers. Since both the vertical and horizontal motions have equal speeds, the slowness term does not come into play. The one layer description fits all the data except for the vertically moving terminators, whereas the two layer description fits all the data perfectly. At low contrast, however, the local likelihoods become much more fuzzy, and therefore the one layer description is favored — even though it does not account for the data perfectly, at low contrasts the fit is good enough to not require an additional model.

3.3.8 The occluded diamond — Lourenceau and Shiffrar 92

Phenomena: Lorenceau and Shiffrar (1992) conducted experiments with the "occluded diamond" shown in figure 3-28. A diamond translates behind invisible occluders. At high contrast, the percept consists of two groups, each line segment moves in the normal direction. Subjects are unable to resolve the "true" motion of the diamond. However, if the stimulus is blurred or viewed peripherally or at low contrast, subjects are capable of perceiving the correct motion.

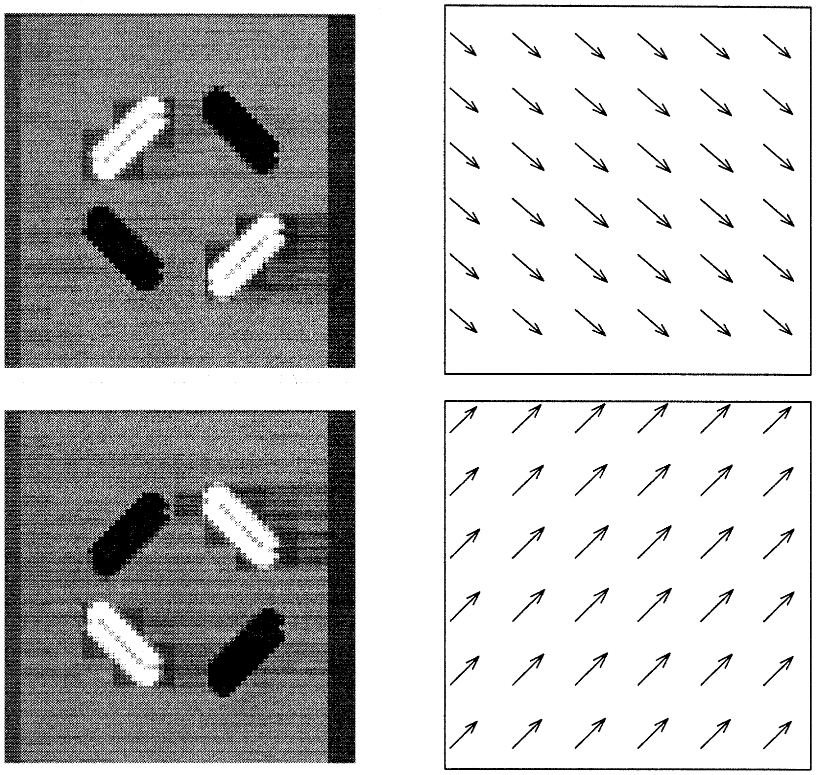


Figure 3-29: Results of the SIL algorithm on the high contrast occluded diamond stimulus. Two groups are found, one corresponding to each pair of diagonal lines

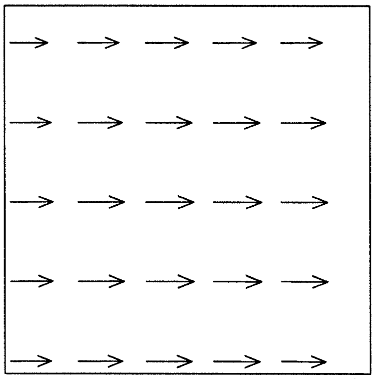


Figure 3-30: Results of the SIL algorithm on the low contrast occluded diamond stimulus. One layer is found, corresponding to the full pattern. Even though the parameters are held constant.

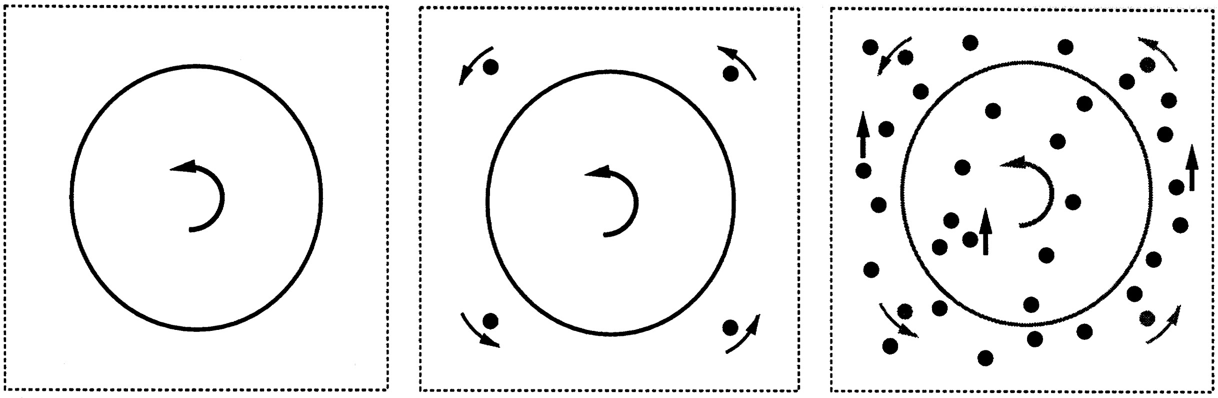


Figure 3-31: When a "fat" ellipse rotates rigidly in the image plane it is perceived as deforming nonrigidly (Wallach et al., 1956). When four rotating dots are added to the display, the ellipse is perceived as rigid (Weiss and Adelson, 1995). The effect of the satellites persists when a large number of vertically translating dots is added to the display (Weiss and Adelson, 1995).

Model Results: Figure 3-29 shows the output of the SIL model at high contrast. Two layers are found, each moving with a diagonal motion. Figure 3-29 shows the output of the model at low contrast. The model finds a single layer moving rightward.

Discussion: This stimulus is nearly identical to the split herringbone and the interpretation is the same. The "failure" to see the correct object motion is actually the most probable solution at high contrast — by using two layers rather than one it is possible to account for all of the data. At low contrast, however, the local likelihoods become much more fuzzy, and therefore the one layer description is favored — even though it does not account for the data perfectly, at low contrasts the fit is good enough to not require an additional model.

3.3.9 Influence of features on percept of rotating ellipse

Phenomena: We have conducted experiments with the rotating ellipse stimulus (Weiss and Adelson, 1995) (see figure 3-31). A "fat" rotating ellipse appears to deform nonrigidly (Musatti, 1924; Wallach et al., 1956; Musatti, 1975). However, when a small number of rotating features are added to the display, the ellipse appears to rotate rigidly. The effect persists when the display is embedded in a field of translating or static dots — the percept consists of two surfaces one containing the ellipse and

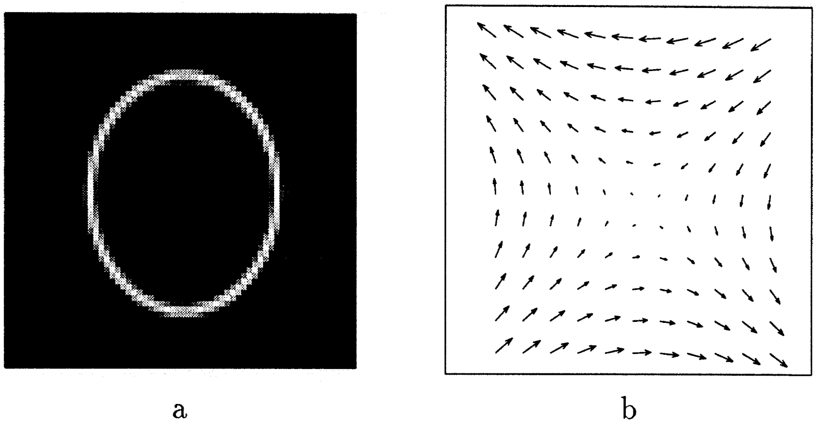


Figure 3-32: **a.** A single frame from a sequence in which an ellipse rotates rigidly in the image plane. **b.** The output of the SIL algorithm. A single layer is found with nonrigid deformation.

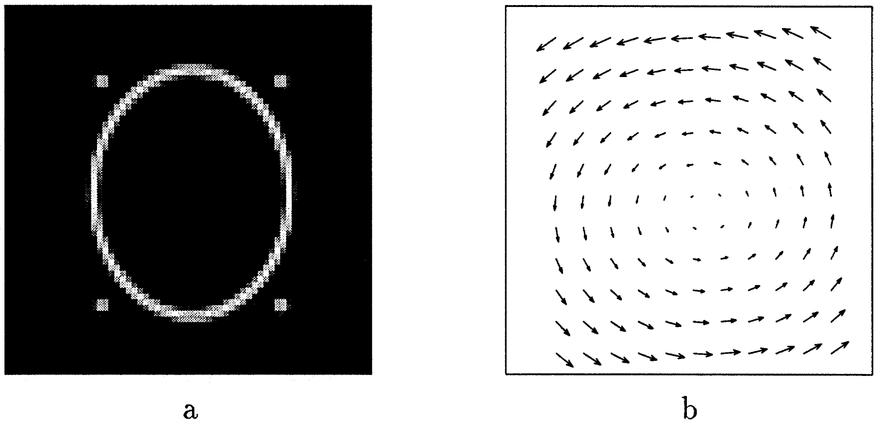


Figure 3-33: **a.** A single frame from a sequence in which an ellipse rotates rigidly in the image plane with four rotating dots. **b.** The output of the SIL algorithm. A single layer is found, with rotational motion.

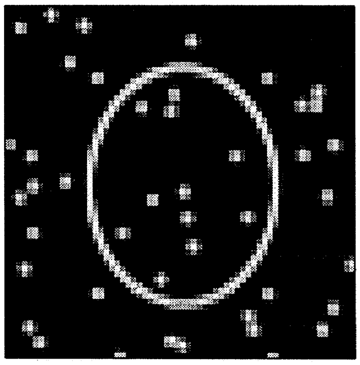


Figure 3-34: A single frame from a sequence in which an ellipse rotates rigidly in the image plane with four rotating dots. An additional 50 dots translate vertically in the image.

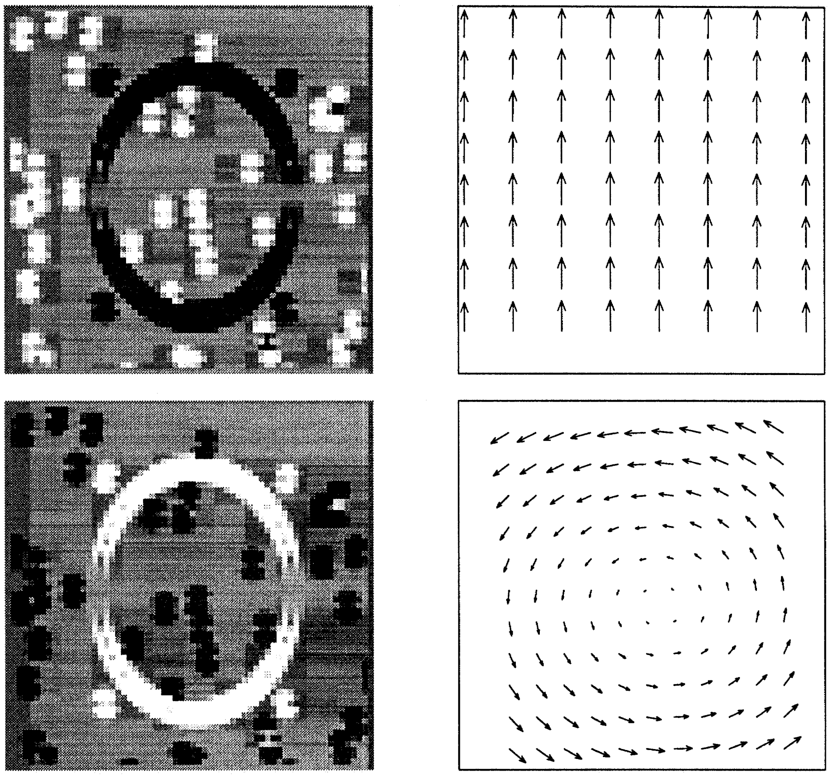


Figure 3-35: The results of the SIL algorithm on the sequence in figure 3-34. Two layers are found, one with rotational motion that includes the ellipse and the four dots, and the other with vertical motion that includes the rest of the dots.

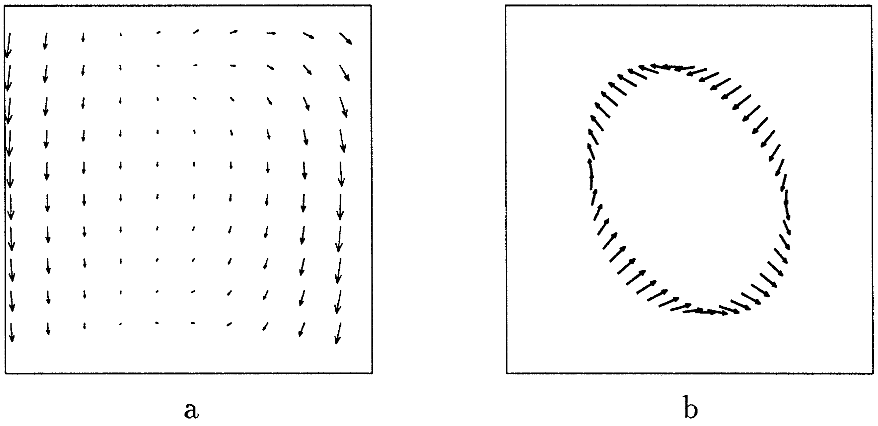


Figure 3-36: **a.** The results of a global smoothness algorithm on the sequence in figure 3-34. The algorithm oversmoothes and the rigid rotation of the ellipse can not be seen. **b.** The result of the Hildreth (1983) algorithm on the the sequence in figure 3-34. The algorithm predicts no interaction between the ellipse's motion and points off the contour. Hence the ellipse is predicted to be nonrigid.

the rotating dots and the other containing the "background" dots.

Model Results: Figure 3-32 shows the output of the SIL model on the plain ellipse stimulus. A single layer is found, deforming nonrigidly. Figure 3-33 shows the output when four rotating dots are added to the display. Again, a single layer is found but now the motion is rotational. Thus the ellipse would be predicted to be perceived as rigid, consistent with human perception. Figure 3-35 shows the output when the scene contains four rotating dots and 50 vertically translating dots. Two layers are found, one corresponding to the ellipse and the four rotating dots, and the other corresponding to the translating background dots. Thus the ellipse would be predicted to be perceived as rigidly rotating, consistent with human perception. The parameters are held constant.

Discussion: As explained in (Weiss and Adelson, 1998) the tendency to see the ellipse by itself as nonrigid is a result of the preference towards slow and smooth motions. When the four dots are added the model prefers a single layer description to the two layers one (in which the ellipse continues to deform slowly and just the dots rotate). The advantage of the two layer description is that the ellipse moves slower, however it requires two layers where a single one suffices. Thus for these parameter settings, the one layer percept is preferred. When the translating dots are added, one would need a very nonsmooth motion to describe the scene using a single layer and the two smooth layers are preferred.

For comparison, figure 3-35a shows the output of a global smoothness algorithm on this sequence. The algorithm oversmoothess and does not predict a rigid rotation for the ellipse. Figure 3-35b shows the output of the Hildreth algorithm on this sequence. The algorithm only combines information along contours and hence there is no influence of the dots on the motion of the ellipse. The predicted ellipse motion is nonrigid.

3.3.10 Simple SIL model — discussion

Our approach throughout this paper has been to search for a small set of assumptions about the world, such that the most probable percept given those assumption would correspond to the percept preferred by humans. The simple SIL model assumes the world consists of a small number of layers, each with a slow and smooth motion field. Unlike the assumption of global smoothness, these assumptions can handle scenes containing transparency and occlusion. Unlike the assumption of smoothness along contours, these assumptions predict an influence of features off the contour on the motion of a contour. Furthermore, we have shown that these assumptions lead to nontrivial predictions about stimuli that will have one motion versus two, and these predictions agree qualitatively with previously published findings.

However, as we noted in the discussion of plaid stimuli, the simple SIL model can not account for all cues that influence the tendency of plaids to cohere. Even for these simple stimuli, static cues to transparency appear to have an influence. In the next section we discuss extensions to the simple SIL model so that form influences can be incorporated.

3.4 Extensions of the simple SIL model

3.4.1 Motivation — some failures of the simple SIL model

A glaring omission from the SIL model is the concept of intrinsic versus extrinsic terminators (Nakayama and Shimojo, 1992). The model assumes that every motion measurement belongs to one of the surfaces. However, there are many measurements in images containing occlusion that are accidental — they are the result of an accidental alignment of two surfaces and their motion is not related to the motion of any of the surfaces. Nakayama and Shimojo (1992) termed these accidental measurements "extrinsic features".

Several researchers have shown that "extrinsic" features tend to have a very week influence on perceived motion (Shimojo et al., 1989; Vallortigara and Bressan, 1991; Ramachandran, 1990). In fact, Anstis (1990) has shown that subjects have great difficulty in tracking these "extrinsic" features when they are instructed to do.

Figure 3-37 shows an example of the weak influence of extrinsic features on mo-

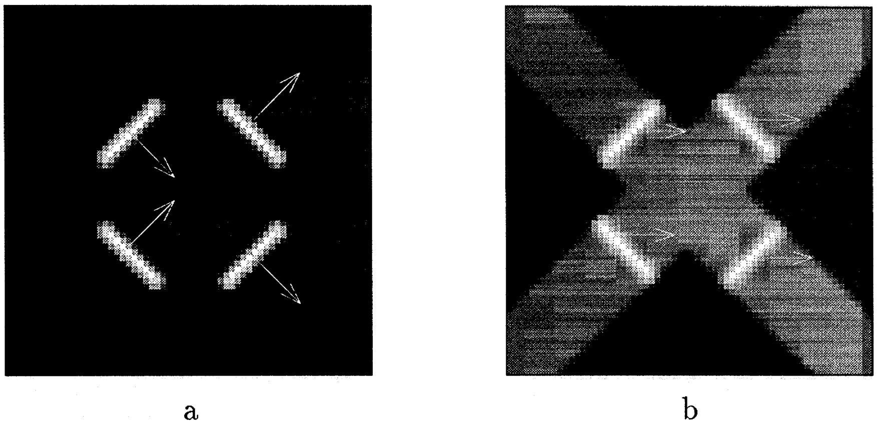


Figure 3-37: The importance of classifying features as extrinsic for motion analysis (Lorenceau and Shiffrar, 1992). **a.** When an occluded diamond translates behind an invisible aperture, two groups are perceived. **b.** When the aperture is made visible, however, subjects tend to perceive the four segments as moving coherently horizontally.

tion perception. Lorenceau and Shiffrar (1992) reported that when the occluders in figure 3-37 were visible, subjects tended to perceive a single coherent diamond. They were not influenced by the "features" caused at the intersection of the occluders and the diamond. These features would be classified as "extrinsic".

What happens when we run the simple SIL algorithm on this display ? We find no effect of the visibility of the occluders. A typical output is shown in figure 3-38. The model finds two layers, each moving with a smooth diagonal motion. The visible aperture is now predicted to move along with the four line segments. This interpretation seems ridiculous to a human observer but it is the most probable one given the simple SIL model. The interpretation favored by humans — where the four line segments move together and the aperture is static, is actually less probable because it does not account for the motion of the endpoints of the line segments. The motion of these endpoints is accidental, but the model has no way of knowing that, nor can it even express the notion of a measurement that belongs to none of the layers.

Static cues not only influence the classification of features into intrinsic or extrinsic, they also come into play when all features are intrinsic. Even after deciding which features to ignore, a visual system still needs to decide which motion measurements

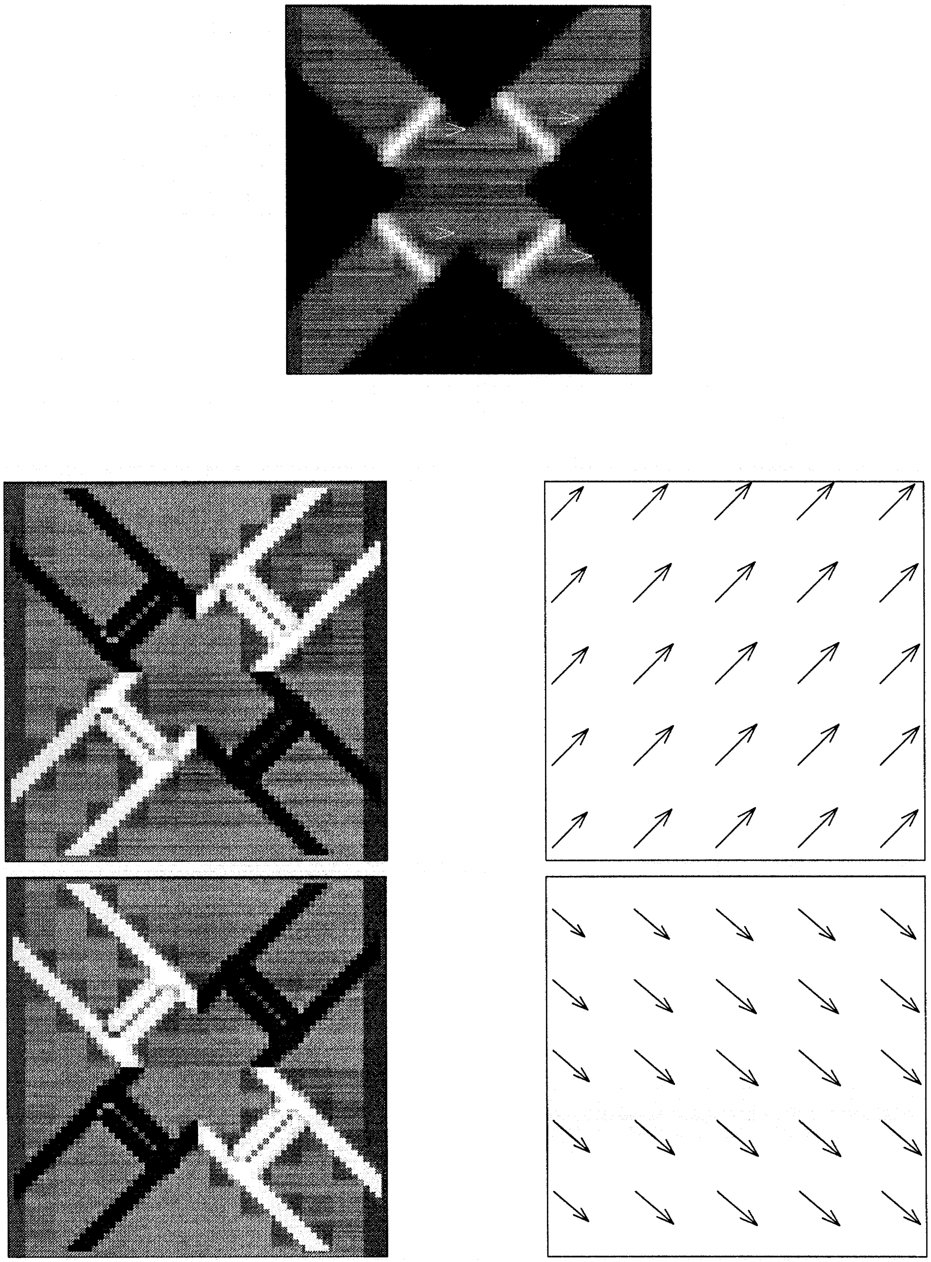


Figure 3-38: The results of running the simple SIL algorithm on the occluded diamond stimulus with a visible aperture. Although humans now tend to perceived the diamond as coherent, the simple model does not do so. Rather it still sees the lines as moving independently and each aperture is predicted to move along its major orientation. This failure of the model is due to the lack of distinction between intrinsic and extrinsic features. Since the model is trying to fit all the motion data, the interpretation that maximizes the posterior probability is "wrong" — it does not correspond to the human percept.

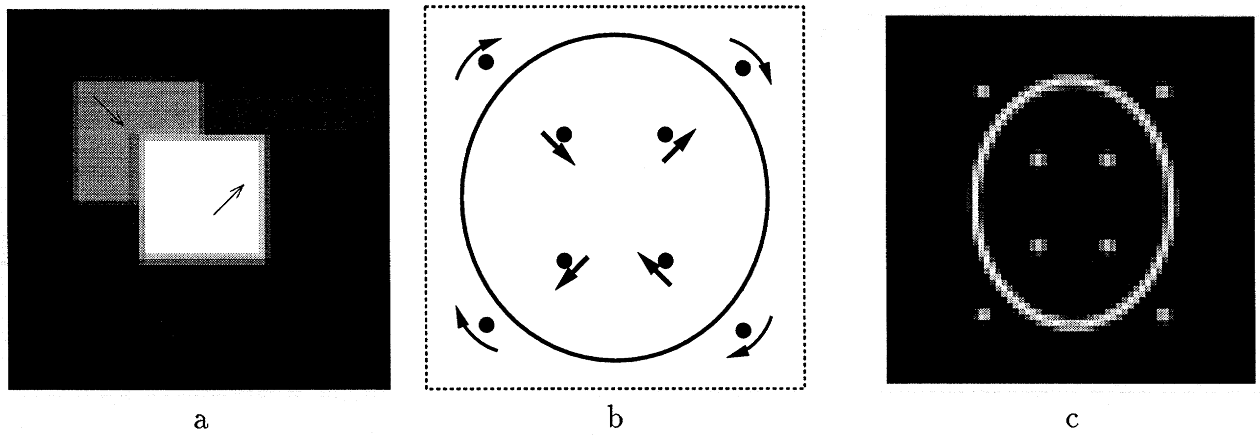


Figure 3-39: Examples of stimuli in which the motion information is not sufficient to determine the assignment into layers, even if extrinsic features are ignored. **a.** A single frame from a sequence in which two squares translate diagonally. **b.** A schematic description of a scene in which an ellipse rotates rigidly in the image plane and two sets of dots move in the image plane. The outer four dots rotate and the inner four dots are consistent with nonrigid deformation. Subjects perceive the ellipse as having the same motion as the closer dots.

go together and which do not. A simple example is shown in figure 3-39 where two squares translate diagonally — one up and to the right and the other down and to the right.

When we run the simple SIL model on this sequence we get the output shown in figure 3-40. The two motions are correctly found but the model does not know how to assign the vertical segments of the squares. These segments are equally well explained by both motions, and yet humans perceiving this scene have no difficulty in determining to which layer they belong. Note that this failure of the simple SIL model has nothing to do with extrinsic/intrinsic classification. In this case, the model is not fooled by the accidental motion of junctions. Rather it does not know how to group together nonaccidental motion signals.

A second example is shown in figure 3-39b. An ellipse rotates rigidly in the image plane and two sets of dots move in the image plane. The outer four dots rotate and the inner four dots are consistent with nonrigid deformation. Subjects perceive the ellipse as having the same motion as the closer dots (Weiss and Adelson, 1995).

The output of the simple SIL algorithm on this sequence is shown in figure 3-41.

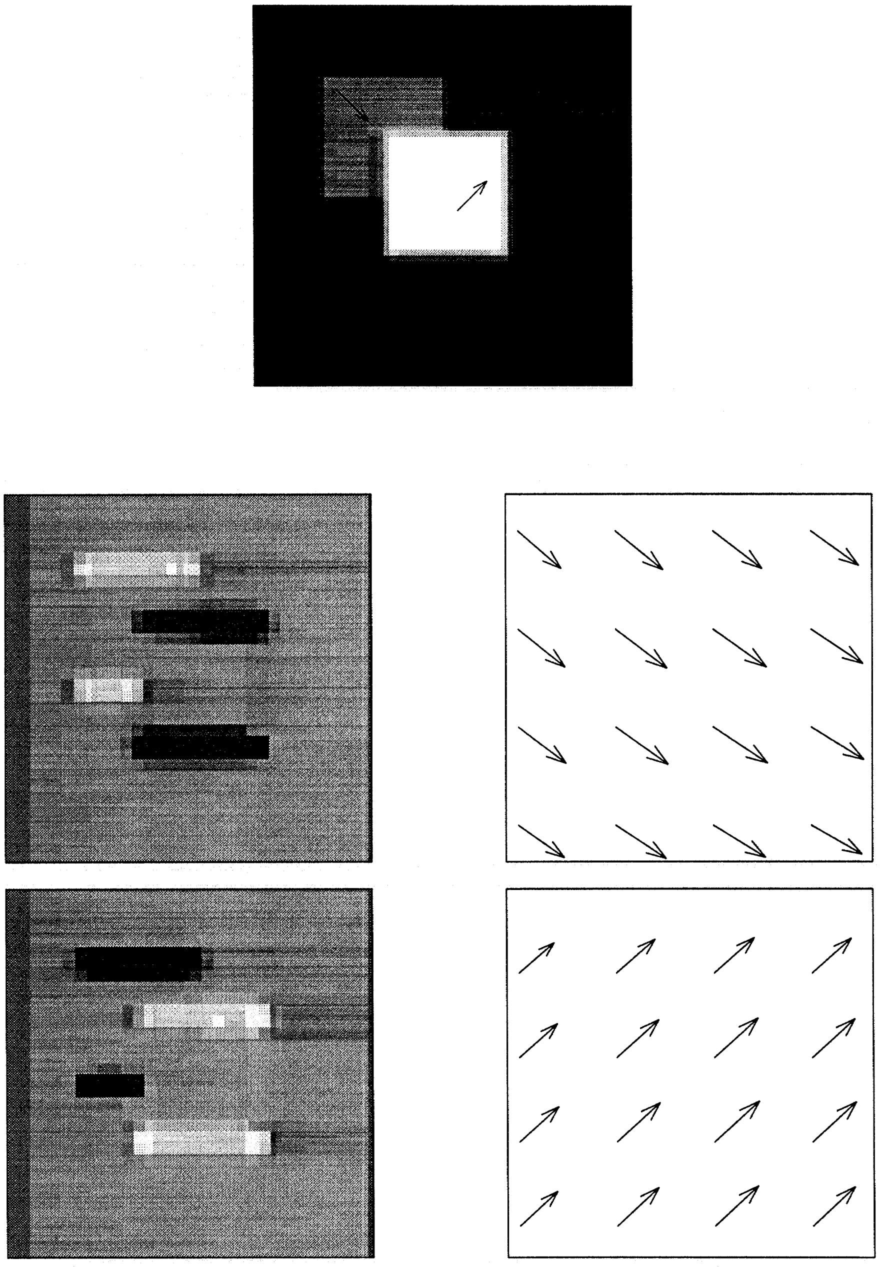


Figure 3-40: The results of running the simple SIL algorithm on the scene with two squares (figure 3-39)a. The simple model correctly estimates the motion of the two layers but does not know how to group the vertical segments of the squares.

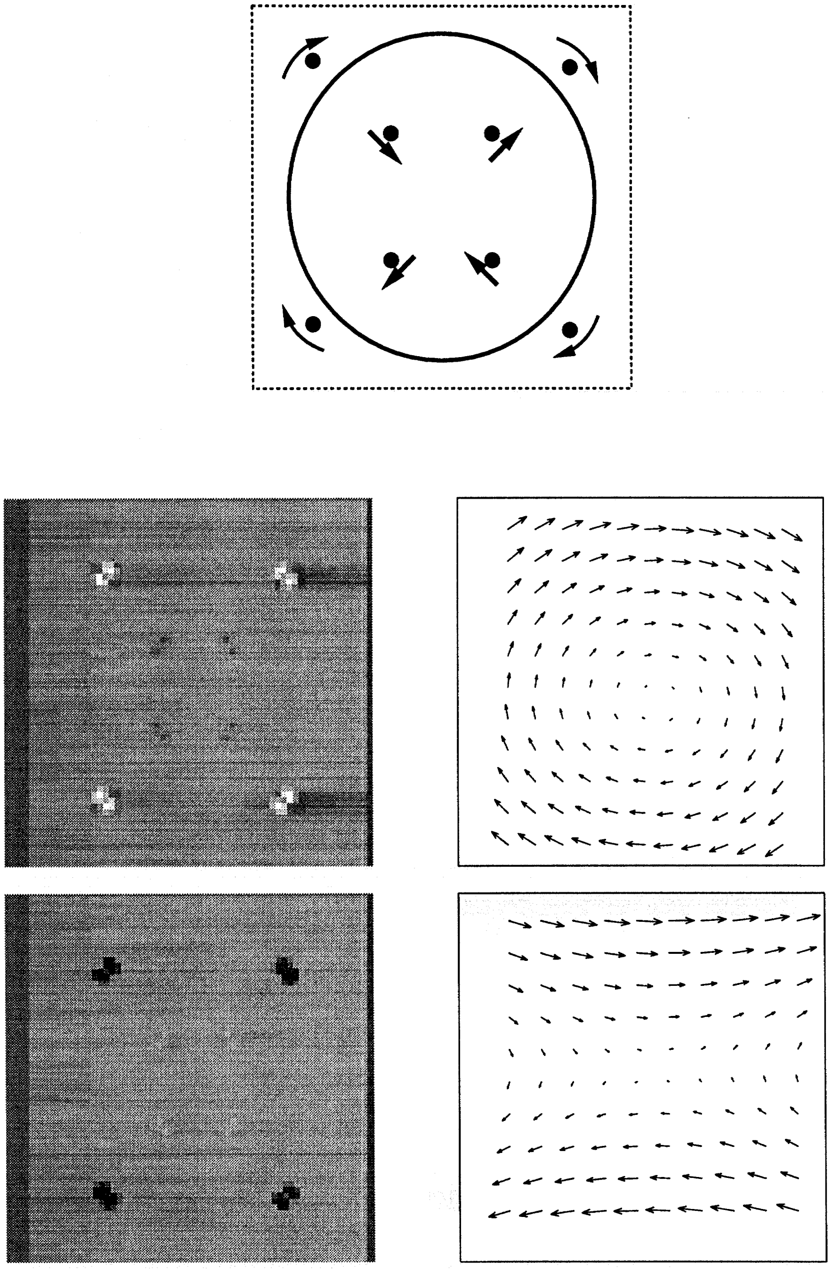


Figure 3-41: The results of running the simple SIL algorithm on the ellipse with two sets of dots. Although subjects see the ellipses as grouped with the the ellipse, the simple model does not do so. Rather it perceives the two groups of dots separately but it does not how to group the ellipse. The measurements along the ellipse's contour are consistent with both motions.

The algorithm correctly finds the motion of the two layers, but it does not know how to group the ellipse. Individual pixels along the ellipse's contour are grouped with one layer or the other depending on local noisy measurements. Again, this is not a failure of the intrinsic/extrinsic distinction but rather an inability to decide which intrinsic features go together.

3.4.2 An extended generative model

In order to incorporate static constraints on the layered decomposition we need to extend the probabilistic model. We present here two extensions to the basic model — the use of outlier processes to model extrinsic features, and the use of Markov Random Field priors on the assignments. The resulting model is still a mixture model, but a slightly less standard one.

Extrinsic features as outliers

Recall that in the generative model behind the simple SIL model, each motion measurement is generated is generated by one of the layers. In order to model extrinsic features, i.e. ones that do not belong to any of the layers, we incorporate an "outlier process" (Black and Rangarajan, 1994) that can also generate data, but does so completely randomly. Outlier processes have been used in previous mixture model formulations (Jepson and Black, 1993) and have an interesting connection to the theory of robust estimation (Black and Rangarajan, 1994). In a typical use of outlier processes in motion segmentation, measurements are classified as "outliers" if they are not consistent with either of the layered motions. The generative model incorporates a prior probability of a measurement being an outlier, and the fact that this number is small prevents all measurements from being classified as outliers.

Unfortunately, this method of adding an outlier process to the mixture model is insufficient for achieving human level performance. While it is true that the motion of extrinsic features will in general not be consistent with the motion of any layer, in some of these displays that is not the case (e.g. figure 3-37). Rather, the junctions are extrinsic because of static cues. Even in the absence of any motion information, it is more likely that T junctions be extrinsic as compared to L junctions. (Nakayama and Shimojo, 1992; Sajda and Finkel, 1994).

To incorporate this into the mixture model, we have a spatially varying prior of being an outlier. That is, rather than having a single, scalar number π that designates the probability of a measurement being an outlier, we have a function $\pi(x, y)$ that designates the probability of a measurement at a particular location being an outlier. This probability is assumed to depend on static form analysis. For example, locations that are adjacent to T-junctions would have a higher probability of being an outlier compared to those adjacent to L-junctions. In the appendix we show how to perform inference on the mixture model with the outlier process and spatially varying priors.

Prior probability on the labeling L

The generative model of the simple SIL model assumes that the labeling of the image is generated randomly. In other words, in the absence of motion information, it is assumed that all assignments of pixels are equally likely. Although this assumptions simplifies the calculations, it is obviously misplaced in the case of motion segmentation. It amounts to the assumption that knowing the membership of a particular location yields no information on the membership of all other locations in the image. In image formation, this is rarely the case: e.g. neighboring points with the same intensity are likely to be from the same object.

In order to model the dependence between labelings at different sites, we add a prior distribution over the labelings $L_k(x, y)$. We use a Markov Random Field (MRF)(Geman and Geman, 1984) distribution:

$$P(L) = \alpha exp \left(\sum_{x,y,x',y'} w_{x,y,x',y'} L^t(x,y) L(x',y') \right) \tag{3.21}$$

The link weights $w_{x,y,x',y'}$ determine the distribution of labelings. For example setting $w_{x,y,x',y'} = 1$ for neighboring sites and zero otherwise makes labelings in which neighboring locations have similar labels more probable.

Inference under this generative model is computationally much more difficult. While an EM algorithm can still be used, the weights $w_{x,y,x',y'}$ make an exact Expectation (or E step) intractable in the general case. Here we use a mean field approximation that has been shown to be relatively successful in other image processing contexts (Zhang et al., 1994). The update rules with the mean field approximation are given in the appendix.

3.4.3 Preliminary Results with the extended SIL model

The problem with the extended SIL model is that it has a very large number of free parameters. The derived segmentations depend on the link weights $w_{x,y,x',y'}$ and the prior probabilities $\pi(x, y)$ that are assumed to be calculated for a given image based on static cues. The static analysis of surface cues is a complicated issue that will necessitate complex models and a full characterization is beyond the scope of this paper. We show here two results to illustrate the potential advantages of using form analysis in this framework.

Figure 3-43 shows the results on the occluded diamond figure when we added the space varying $\pi(x, y)$. To set the prior probabilities for being an outlier, we first ran a primitive "T-junction" detector on the image. The detector searches for regions containing a trimodal intensity distribution. The output of the T-junction detector is shown in figure 3-42 superimposed on the original image. Locations with added white are those where a T-junction has been found. Although the primitive T-junction detector gives reasonable answers for this particular image, it is by no means a general purpose junction finder. More sophisticated junction finders are described in (Freeman, 1992).

In the locations in which T-junctions were found we set $\pi(x, y) = 1$, i.e. these locations are definitely outliers. The results shown in figure 3-43 are obtained with the same parameters σ_N, σ_P as in figure 3-38 but now the solution that maximizes the posterior probability is the one perceived by humans. The model does not attempt to fit the motions of the terminators because it classifies them as outliers. The diamond is therefore predicted to move coherently, consistent with the results of human observers.

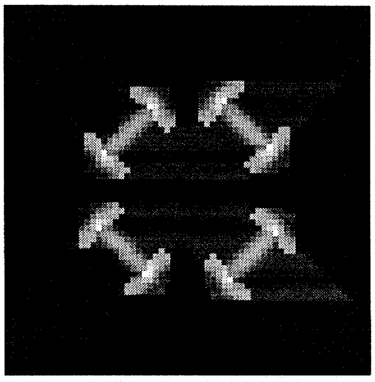


Figure 3-42: The results of running a "t-junction" detector on the image in 3-37b (the output of the detector is superimposed on the original image). We used a very rudimentary "t-junction" detector here that searches for regions with a tri-modal intensity distribution. More sophisticated and biologically plausible detectors have been proposed elsewhere (e.g. (Freeman, 1992)). Our goal here is to show how the outputs of such a detector could influence the motion grouping.

In our formulation terminators are weighted in accordance with their probability of being extrinsic. That is, the formulation allows a gradual distinction between intrinsic and extrinsic features, rather than a sharp classification. Thus in the classic barberpole illusion (Wallach, 1935), the terminators are typically T-junctions and yet they have an influence on the perceived direction of motion. Rubin and Hochstein (1992) (Rubin and Hochstein, 1993) have shown that the influence of terminators changes gradually as more information is available that they are occlusion related. In our framework this would be modeled as an increasing probability of being an outlier.

We again emphasize that the simple junction finder used here is not meant to give a full model of the interaction between surface cues and the intrinsic/extrinsic distinction. Rather we wish to illustrate the capability of the mixture framework for handling these additional cues. For example, the results of Stoner et al. (1990) suggest that a particular type of X junction that is more consistent with transparency is more likely to be perceived as extrinsic. This would of course not "fall out" of the simple junction finder we use here. Rather it would have to be put into the model specifically. The advantage of the extended framework is that it can incorporate form cues for being an outlier. It does not, however, prescribe a method for finding the form cues.

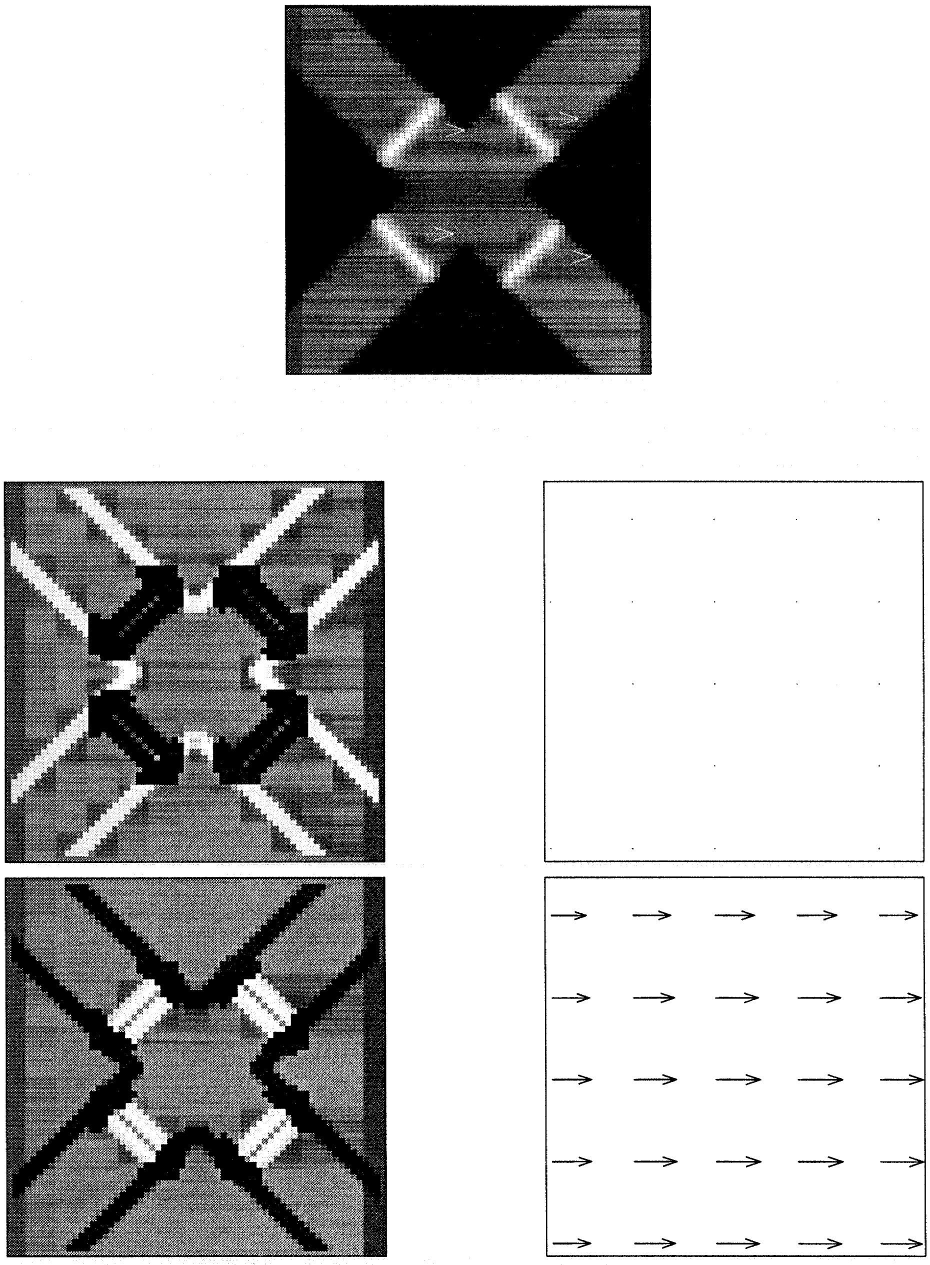


Figure 3-43: The results of running the SIL algorithm with spatially varying outlier priors on the occluded diamond stimulus with a visible aperture. The output of the T-junction detector was used to identify measurements that are likely to be outliers. The parameters σ_N, σ_P are identical to those used in figure 3-38 but the interpretation that maximizes the posterior probability is now "correct" – i.e. it corresponds to human perception of these displays.

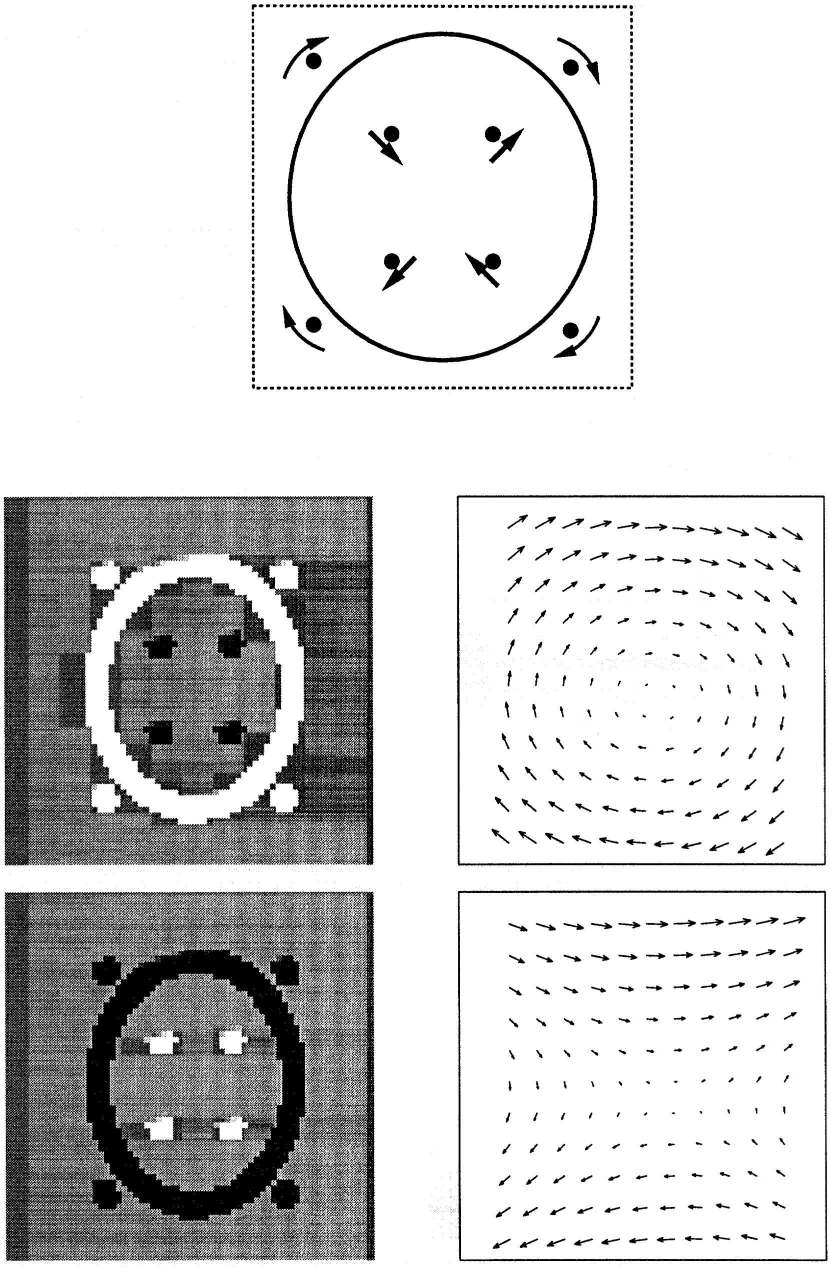


Figure 3-44: The results of running the SIL algorithm with "proximity" priors on the labelings on the stimulus in figure 3-39b. The parameters σ_N, σ_P are identical to those used in figure 3-41 but the interpretation that maximizes the posterior probability is now "correct" – i.e. it corresponds to human perception of these displays. Because the priors favor groupings that obey proximity, the ellipse is more likely to be grouped with the set of dots that is closer to it.

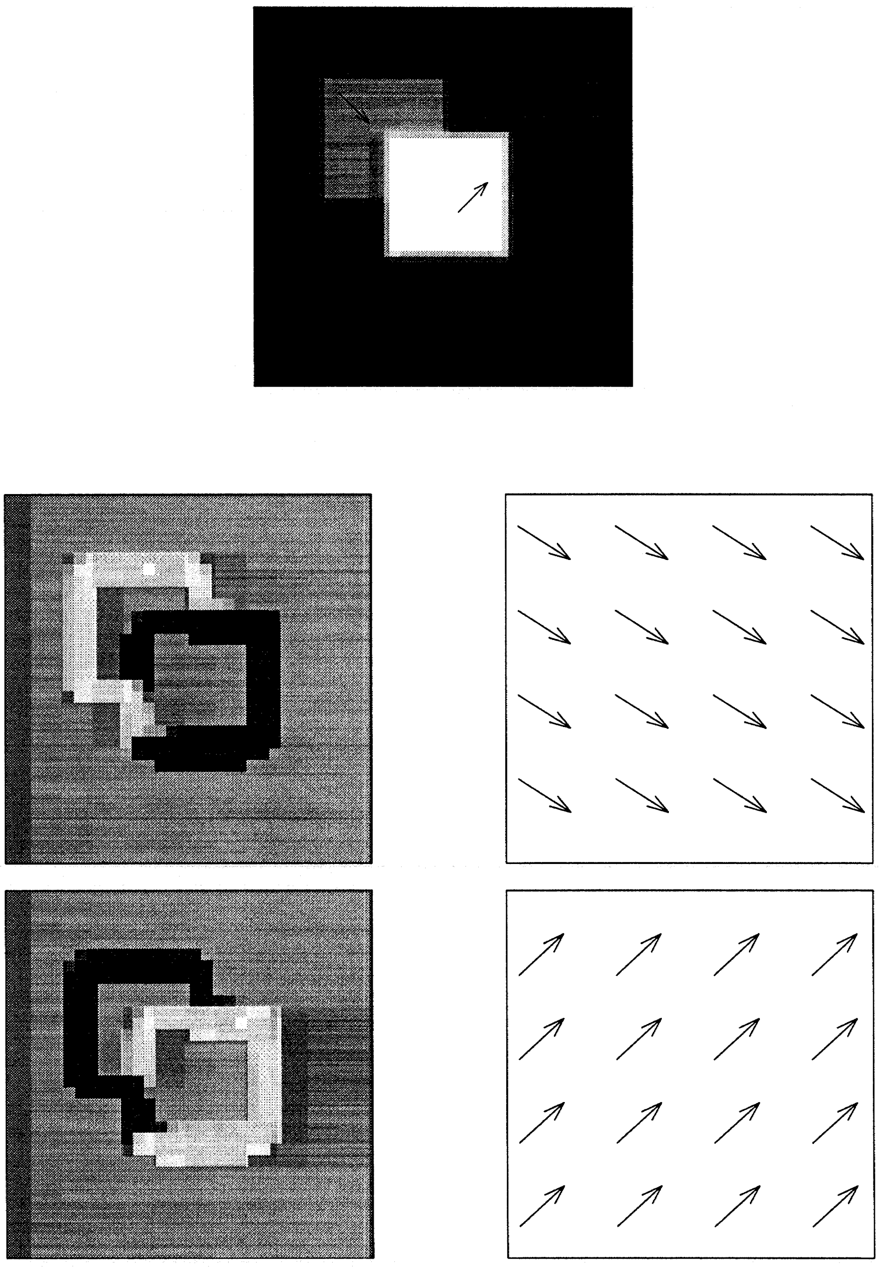


Figure 3-45: The results of running the SIL algorithm with "proximity" priors on the labelings on the stimulus in figure 3-39a. The parameters σ_N, σ_P are identical to those used in figure 3-40 but the vertical segments are now assigned to one layer or the other based on proximity. Note that the assignment is still not completely correct, suggesting that more sophisticated priors may be needed.

Figure 3-44 shows the output of the model when proximity priors are used on the labelings. The link weights $w_{x,y,x',y'}$ were set so they favored proximity:

$$w_{x,y,x',y'} = e^{-\frac{(x-x')^2+(y-y')^2}{2\sigma_w^2}} \tag{3.22}$$

with the added constraint that $w_{x,y,x',y'}$ was set to zero for locations that had no motion information (i.e. the uniform background).

In this case, the ellipse is grouped with the set of dots that is closer to it, consistent with the human percept. Evidence that the "proximity" prior is not enough, however, is shown in figure 3-45. This is the output obtained by the model on two squares scene of figure 3-39a. Note that the assignment is still not completely correct. One of the vertical segments is accidentally grouped with the wrong square. In addition, the insides of the squares are not assigned to either square. This suggests the need for more sophisticated link weights w, e.g. incorporating "good continuation" or "figure-ground".

3.4.4 Discussion — extended SIL model

The preliminary results presented here illustrate both the advantages and the disadvantages of the extended SIL model. Unlike the simple SIL model, the extended one can incorporate static form constraints to perform the classification of terminators into extrinsic/intrinsic and the grouping together of intrinsic measurements. However, in order to do so, the extended model relies on static form analysis and has a large number of degrees of freedom — what type of junction detector is used, the numerical value given to the probability of a particular junction being extrinsic, the link weights between various locations etc. Thus unlike the simple SIL model, the extended model by itself does not give strong, nontrivial predictions. It remains a challenge for the future to see whether a small number of principles can be incorporated into the extended SIL model so that they will explain a range of percepts.

3.5 Discussion

Estimating motion in scenes containing multiple motions imposes conflicting demands on the visual system (Braddick, 1993). In order to overcome the local ambiguity of the individual measurements, the system must combine information across space. At the same time, however, the system needs to avoid mixing together measurements from different objects. In this paper we have proposed a computational framework for this simultaneous estimation and segmentation, a framework we call "smoothness in layers". The model is primarily motivated by the inability of existing models to handle basic percepts involving occlusion, transparency and motion capture.

Global smoothness algorithms can successfully account for human perception in scenes containing a single motion. For scenes containing occlusion and transparency, however, they predict a single nonrigid motion in cases where humans perceive multiple smooth motion fields. Algorithms that only assume smoothness along contours are more successful at handling occlusion, but they predict no interaction between features off the contour and the motion of the contour. The model presented here can account for this range of percepts in a single framework.

The framework also provides a context in which to investigate the assumptions that are used by the visual system in order to derive a surface representation from the retinal input. We have shown that a simple model with a small number of assumptions can account for a range of phenomena including the tendency of gratings to cohere into plaids and the influence of dots off the contour on the motion of a smooth contour. The simple model attempts to describe the motion data with a small number of layers, each with a slow and smooth motion field. We have also shown stimuli that can not be accounted for with these small set of assumptions and suggested how to incorporate additional assumptions into the model.

The connection between perceived coherence of plaids and the tendency towards slow speeds has been previously suggested in (Farid et al., 1995; Simoncelli, 1993). Using a distributed model for local velocity, they suggested that plaids appear incoherent when the distributed representation of velocity tuned units is multimodal (Simoncelli, 1993; Simoncelli and Heeger, 1998). Bimodality as a corelate of perceptual transparency has also been suggested in (Wilson and Kim, 1994; Jasinschi et al., 1992). Although these models can each account for a subset of the phenomena considered here, it cannot account for the more complicated motions such as the capture of smooth curves undergoing rotation by a small number of dots. Here we have shown how a single concept, smoothness in layers, can account for transparency in plaids, occluded diamonds and rotating dots and ellipses.

A distributed representation of velocity was also the basis of the model of Nowlan and Sejnowski (1995) . However, unlike the models surveyed above, their model included two subpopulations of motion selective units: (a) velocity tuned units similar to those used in (Simoncelli and Heeger, 1998) and (b) selection units. The selection units calculate an estimate of confidence in the local velocity estimate, i.e. based on the local spatiotemporal filters they estimate how good the evidence is for every velocity. The selection units are trained and "become sensitive to regions containing motion energy at several orientations, since velocity predictions from these regions contain sufficient information to disambiguate the direction and speed of true motion". They suggest that the activity of the selection units may be helpful in a later segmentation stage, but their model does not explicitly decide which motions are present in the scene or the assignment of locations to motions.

Shizawa and Mase (1991) suggested an alternative method for determining the number of motions in a local region. They reformulated the standard local optical flow constraints to allow for the input image being a linear superposition of two images, each moving in a different direction. Their algorithm chooses a single motion whenever the information can be explained with a single motion, and hence, would never predicts incoherence in plaids.

Another interesting distinction between the Shizawa and Mase (1991) model and many layered models (including our own) is that it actually assumes every motion measurement was generated by a superposition of the two motions. In contrast, our model assumes every datapoint belongs to one model or the other. Thus when we segment an additive transparency stimulus (such as the plaids) we actually assign individual locations to one grating or the other. For our segmentation algorithm to work, there need to be individual measurements that are predominantly due to one surface or the other. Interestingly, Qian et al (1994) have found that displays that appear transparent "always contain locally unbalanced motion signals, with some local regions having net motion signals in one direction and some other regions in the opposite direction". However, the notion of "locality" used by Qiyan et al. includes locality in spatial frequency and orientation. We require locality in space and hence our inability to account for transparency in sine-wave plaids.

Layered motion models have become increasingly popular in computer vision (Darrell and Pentland, 1991; Jepson and Black, 1993; Irani and Peleg, 1992; Hsu et al., 1994; Ayer and Sawhney, 1995; Wang and Adelson, 1994). The model described here shares many of the features of these algorithms but they needed to be modified in order to account for human performance. First, these models typically assume that each layer is moving rigidly in $3D$ but humans can perceive nonrigid motion groups (e.g. the deforming ellipse with nonrigid dots). Hence the assumption of smoothness in layers. Second, most of these algorithms assume the number of layers is given in advance (although see (Ayer and Sawhney, 1995) for an exception). In order to account for human perception, the interesting question is when humans see one motion rather than two. To the best of our knowledge, none of the layered motion models cited above would predict incoherence in plaid patterns, to say nothing about the change from coherence to incoherence when the speed is varied.

The method we propose here for estimating the number of layers is by no means the only possible method. Alternative model selection methods include "minimum description length (MDL)", "minimum message length" (MML), "Akaiake information criterion" (AIC) and "Bayesian information criterion" (BIC) (see (Torr, 1998) for a recent review). These methods choose an interpretation that maximizes the sum of the posterior probability and a complexity term that rewards simple descriptions. For example, in the case of plaids, while we choose between the coherent and the transparent description based on which maximizes the posterior probability, the alternative methods would choose the description that maximizes the posterior plus the complexity term. In all these methods, however, the complexity term does not depend on the motion data and hence would not change when stimulus attributes are changed. Thus adding a complexity term would not change the pattern of results reported here — it would shift the absolute threshold where the percept shifts from transparent to coherent but the trends would remain identical.

From a computational standpoint, the smoothness in layers model is similar to that proposed by Marroquin (1992) and by Madrasmi et al. (1993). They dealt with the more general problem of surface segmentation and proposed a relaxation network that estimates multiple, smooth surfaces by assuming smoothness within each surface. The optimization was performed using simulated annealing. The main computational difference between that work and our own is the introduction of the mixture model in our case, that enables the estimation of the number of layers automatically in a probabilistic framework. The probabilistic framework is particularly important when dealing with motion measurements that have varying degrees of ambiguities.

Shizawa (1993) also considered the general problem of multiple transparent surface estimation. He showed how to avoid the nonlinear optimization performed in (Marroquin, 1992; Madrasmi et al., 1993) by a clever reformulation of the problem. This leads to an approximation to the cost function that is linear in its parameters, can be optimized much more quickly and allows for solving for the multiple surface values without an explicit assignment of points to surfaces. Unfortunately, the approximation leads to erroneous results whenever the two surfaces intersect. As a framework for modeling human visual performance this approach also suffers from its inability to estimate the number of surfaces. Furthermore, it does not explicitly represent the presence of multiple surfaces in the scene. Rather there are multiple values at every location, but the grouping of these values into global surfaces is left for further processing.

As mentioned in the introduction, the theory presented here is at the computational level in the categorization described by Marr (1982) — we make no claim about the algorithm or the neuronal implementation. Given the complexity of maximizing the likelihood in a mixture model, it seems hard to fathom how effortless the perception of motions in scenes containing multiple motions is. Somehow, humans are able to estimate the number of surfaces, the belongingness of locations into layers and the motion of each layer. This suggests that the brain has found a way to solve the same problems solved in mixture estimation, but in a remarkably fast and robust fashion.

3.6 Appendix

3.6.1 EM update equations for line fitting

For completeness we give here the EM update equations for line fitting. Recall that we are given $\{x_i, y_i\}$ and wish to estimate $\{a_k, b_k\}$ the parameters of the lines.

E step: Define $R_k^2(i) = (a_k x_i + b_k - y_i)^2$ the squared residual between the predicted $y(x_i)$ and the actual y_i. Compute the posterior probability:

$$\hat{L}_k(i) = \frac{e^{-R_k^2/2\sigma^2}}{\sum_j e^{-R_j^2/2\sigma^2}} \tag{3.23}$$

M step: For a given line k, minimize:

$$J(a_k, b_k) = \sum_i L_k(i)(a_k x_i + b_k - y_i)^2 \tag{3.24}$$

Thus denoting the vector $\vec{\theta_k} = (a_k, b_k)^T$, $\vec{\theta}$ is a solution to $M\theta_k = b_k$ with:

$$M = \sum_i \hat{L}_k(i) \begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & \sum_i 1 \end{pmatrix} \tag{3.25}$$

and:

$$b = \sum_i \hat{L}_k(i) \begin{bmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{bmatrix} \tag{3.26}$$

3.6.2 EM iterations for curve fitting

Here we estimate arbitrary smooth curves rather than a low dimensional parametric curve. We assume a prior distribution over curves:

$$P(y(x)) = \alpha e^{-J(y)/\sigma_P^2} \tag{3.27}$$

with:

$$J(y) = \int \|Dy(x)\|^2 dx \tag{3.28}$$

We use a smoothness functional such that the optimal curve will be given by a superposition of Gaussians (Girosi et al., 1995) $y(x) = \sum_j \theta_j G(x - x_j)$. We are given $\{x_i, y_i\}$ and wish to estimate $\{\theta^k\}$ the parameters of the curves. Note that θ^k is a vector whose length is equal to the number of datapoints.

E step: Define $R_k^2(i) = (\sum_j \theta_j^k G(x_i - x_j) - y_i)^2$ the squared residual between the predicted $y(x_i)$ and the actual y_i. Compute the posterior probability:

$$\hat{L}_k(i) = \frac{e^{-R_k^2/2\sigma_N^2}}{\sum_j e^{-R_j^2/2\sigma_N^2}} \tag{3.29}$$

M step: For a given curve k, minimize:

$$J(\theta^k) = \sum_i \hat{L}_k(i) R_k^2(i) + \frac{\sigma_N^2}{\sigma_P^2} \theta R \theta \tag{3.30}$$

with $R_{ij} = G(x_i - x_j)$. Taking the derivative with respect to θ^k gives the following system of equations:

$$(WR + \frac{\sigma_N^2}{\sigma_P^2} I)\theta = WY \tag{3.31}$$

where W is a diagonal matrix whose entries are $W_{ii} = \hat{L}_k(i)$ and Y is a vector whose entries are y_i.

3.6.3 EM Iterations for smoothness in layers

Here we estimate a smooth velocity field for every layer. We are given I_x, I_y, I_t the spatial and temporal derivatives at every locations. Recall that we parameterize every velocity field with a 50 dimensional vector:

$$v_x^k(x, y) = \sum_{i=1}^{25} \theta_i^k G(x - x_i, y - y_i) \tag{3.32}$$

$$v_y^k(x, y) = \sum_{i=26}^{50} \theta_i^k G(x - x_i, y - y_i) \tag{3.33}$$

where $G(x, y)$ is a Gaussian in image space and $\{x_i, y_i\}$ form a $5x5$ grid over the image. Throughout this paper we used Gaussian basis functions whose standard deviation was 2/3 the size of the image. We assume a prior distribution over velocity fields:

$$P(V) = \alpha e^{-J(V)/2\sigma_P^2} \tag{3.34}$$

with:

$$J(V) = \sum_{xy} \|Dv(x,y)\|^2 \tag{3.35}$$

here Dv is a differential operator, i.e. it measures the derivatives of the velocity field. We follow Grzywacz and Yuille (1991) in using a differential operator that penalizes velocity fields with strong derivatives:

$$Dv = \sum_{n=0}^{\infty} a_n \frac{\partial^n}{\partial x} v \tag{3.36}$$

We again chose the differential operator so that the Gaussian basis functions would be the Green's functions for this operator. Recall also that we use a local likelihood for a velocity at a pixel defined by:

$$P(I_x, I_y, I_t | v_x, v_y) = \alpha e^{-C(v_x, v_y)/2\sigma_N^2} \tag{3.37}$$

where $C(x,y)$ quantifies the degree of consistency of the velocity with the local data. It is based on the gradient constraint (Horn and Schunck, 1981; Lucas and Kanade, 1981):

$$C(v_x, v_y) = \sum_{x,y,t} w(x,y,t)(I_x v_x + I_y v_y + I_t)^2 \tag{3.38}$$

where v_x, v_y denote the horizontal and vertical components of the local velocity I_x, I_y, I_t denote the spatial and temporal derivatives of the intensity function and $w(x,y,t)$ is a spatiotemporal window centered at (x,y,t).

E step: Given θ^k, define $C_k^2(x,y)$ by substituting the local velocities in equations 3.32 into equation 3.38. Compute the posterior probability:

$$\hat{L}_k(x,y) = \frac{e^{-C_k^2/2\sigma_N^2}}{\sum_j e^{-C_j^2/2\sigma_N^2}} \tag{3.39}$$

M step: The M step is identical to the equations detailed in (Weiss and Adelson, 1998) except that the local likelihoods are weighted by $\hat{L}_k(x,y)$. To simplify the notation, we denote the location (x,y) with a single vector r. We define a matrix $\Psi(r)$ whose components give the values of the basis fields at location r:

$$\Psi_{ij}(r) = G(r - r_j) \tag{3.40}$$

So that $v(r) = \Psi(r)\theta$.

By completing the square, the local likelihood at location r can be rewritten:

$$\begin{aligned} P(I_x, I_y, I_t | v_x, v_y) &= \alpha e^{-C(v_x,v_y)/2\sigma_N^2} && (3.41) \\ &= \alpha e^{-(\Psi(r)\theta - \mu(r))^t \Sigma^{-1}(r)(\Psi(r)\theta - \mu(r))/2\sigma_N^2} && (3.42) \end{aligned}$$

To find θ_k we we solve:

$$A\theta^* = b \tag{3.43}$$

with:

$$A = \left(\sum_r \hat{L}_k(r) \Psi^t(r) \Sigma^{-1}(r) \Psi(r)/\sigma^2 + R/\sigma_p^2 \right) \tag{3.44}$$

$$b = \left(\sum_r \hat{L}_k(r) \Psi^t(r) \Sigma^{-1} \mu(r) \right) / \sigma^2 \tag{3.45}$$

and R the prior matrix defined as in (Weiss and Adelson, 1998). If $i <= 25, j <= 25$ or if $i > 25, j > 25$ then $R_{ij} = G(x_i - x_j, y_i - y_j)$. Otherwise, $R_{ij} = 0$.

3.6.4 Posterior probability for smoothness in layers

Since EM may converge to a local maximum of the posterior, we use multiple restarts of the algorithm and choose the output that has highest posterior probability.

Given a set of candidate parameter vectors Θ, we first calculate the velocity fields (equation 3.32) and that in turn gives us $C_k^2(r)$ (equation 3.38). The mixture log likelihood is:

$$l = \sum_r \log(\sum_k e^{-C_k^2/2\sigma_N^2}) \tag{3.46}$$

The log posterior can be written:

$$\log(P(\Theta)) = -\frac{1}{2\sigma_P^2} \sum_k \theta_k^t R \theta_k \tag{3.47}$$

and the sum of the log posterior and the log likelihood gives the log posterior.

3.6.5 EM update rules for smoothness in layers with outlier process

Here we assume that there are K layers plus an additional outlier process. For every location (x, y) we have a prior probability $\pi_k(x, y)$ that this location belongs to one of the layers or the outlier process. The M step is unchanged, only the E step changes.

The E step now becomes:

$$\hat{L}_k(x, y) = \frac{\pi_k(x, y) e^{-C_k^2/2\sigma_N^2}}{\sum_j \pi_j(x, y) e^{-C_j^2/2\sigma_N^2}} \tag{3.48}$$

where C_{K+1}, the consistency of the outlier process is a constant. We used $C_{K+1} = 3\sigma_N^2$.

As mentioned in the paper, $\pi_k(x, y)$ is set based on form processing. For the examples in this paper, we first ran the primitive "T-junction" detector on the image. For locations in which a "T-junction" was found we set $\pi_{K+1}(x, y) = 1$ and $\pi_k(x, y) = 0$ for all $k <= K$. For locations in which no "T-junction" was found we set $\pi_{K+1}(x, y) = 0$ and $\pi_k(x, y) = 1/k$ for all $k <= K$.

3.6.6 EM with MRF priors and mean field approximation

In this case we are given a set of "link weights": w_{rs} that specify a prior probability over weightings.

$$P(L) = \alpha e^{\sum_{rs} w_{rs} L^t(r) L(s)} \tag{3.49}$$

Again the M step does not change. An exact E step is intractable but a commonly used approximation is the mean field approximation. The E step under the mean field approximation calls for iteratively updating $\hat{L}_k(r)$ by first collecting "votes" $V_k(r)$:

$$V_k(r) = \sum_s w_{rs} \hat{L}_k(s) \tag{3.50}$$

and then updating $L_k(r)$ by:

$$\hat{L}_k(r) = \frac{e^{-C_k^2/2\sigma_N^2} - V_k}{\sum_j e^{-C_j^2/2\sigma_N^2 - V_j}} \tag{3.51}$$

We used a serial update schedule in which we visited all pixels once per E step in scanline order.

3.6.7 Calculating the posterior probability for an ideal square wave plaid

Consider a square wave plaid consisting of ideal lines at orientations θ_1, θ_2, contrast C and period $1/p$ moving with velocity v. We assume the length of visible lines is given by L. The spatio temporal derivatives are Cn_1 and Cn_2 where n_1 is a unit vector whose orientation is perpendicular to θ_1 and n_2 is perpendicular to θ_2. The temporal derivatives are $y_1 = Cv^t n_1$ and $y_2 = Cv^t n_2$.

The likelihood (equation 3.46) has a particularly simple form for these sequences, as there are only three types of terms:

$$\begin{aligned} l &= \sum_r \log(\sum_k e^{-C_k^2/2\sigma_N^2}) && (3.52)\\ &= \alpha_1 \log(\sum_k e^{-D_k^2/2\sigma_N^2}) && (3.53)\\ &\quad +\alpha_2 \log(\sum_k e^{-E_k^2/2\sigma_N^2}) && (3.54)\\ &\quad +\alpha_3 \log(\sum_k e^{-F_k^2/2\sigma_N^2}) && (3.55) \end{aligned}$$

where $\alpha_1 = \alpha_2 = Lp$ and $\alpha_3 = p^2$, and:

$$D_k^2 = (n_1 v^k + y_1)^2 \tag{3.56}$$

$$E_k^2 = (n_2 v^k + y_2)^2 \tag{3.57}$$

$$F_k^3 = (n_1 v^k + y_1)^2 + (n_2 v^k + y_2)^2 \tag{3.58}$$

where v^k is the velocity of layer k.

If we are considering translational velocity fields for the different layers then the prior also has a simple form:

$$\log(P(v^1, v^2)) = -\|v^1\|^2/2\sigma_P^2 - \|v^2\|2/2\sigma_P^2 \tag{3.59}$$

Chapter 4

Smoothness in Layers: Motion segmentation using nonparametric mixture estimation

Abstract

Grouping based on common motion, or "common fate" provides a powerful cue for segmenting image sequences. Recently a number of algorithms have been developed that successfully perform motion segmentation by assuming that the motion of each group can be described by a low dimensional parametric model (e.g. affine). Typically the assumption is that motion segments correspond to planar patches in 3D undergoing rigid motion. Here we develop an alternative approach, where the motion of each group is described by a smooth dense flow field and the stability of the estimation is ensured by means of a prior distribution on the class of flow fields. We present a variant of the EM algorithm that can segment image sequences by fitting multiple smooth flow fields to the spatiotemporal data. Using the method of Green's functions, we show how the estimation of a single smooth flow field can be performed in closed form, thus making the multiple model estimation computationally feasible. Furthermore, the number of models is estimated automatically using similar methods to those used in the parametric approach. We illustrate the algorithm's performance on synthetic and real image sequences.

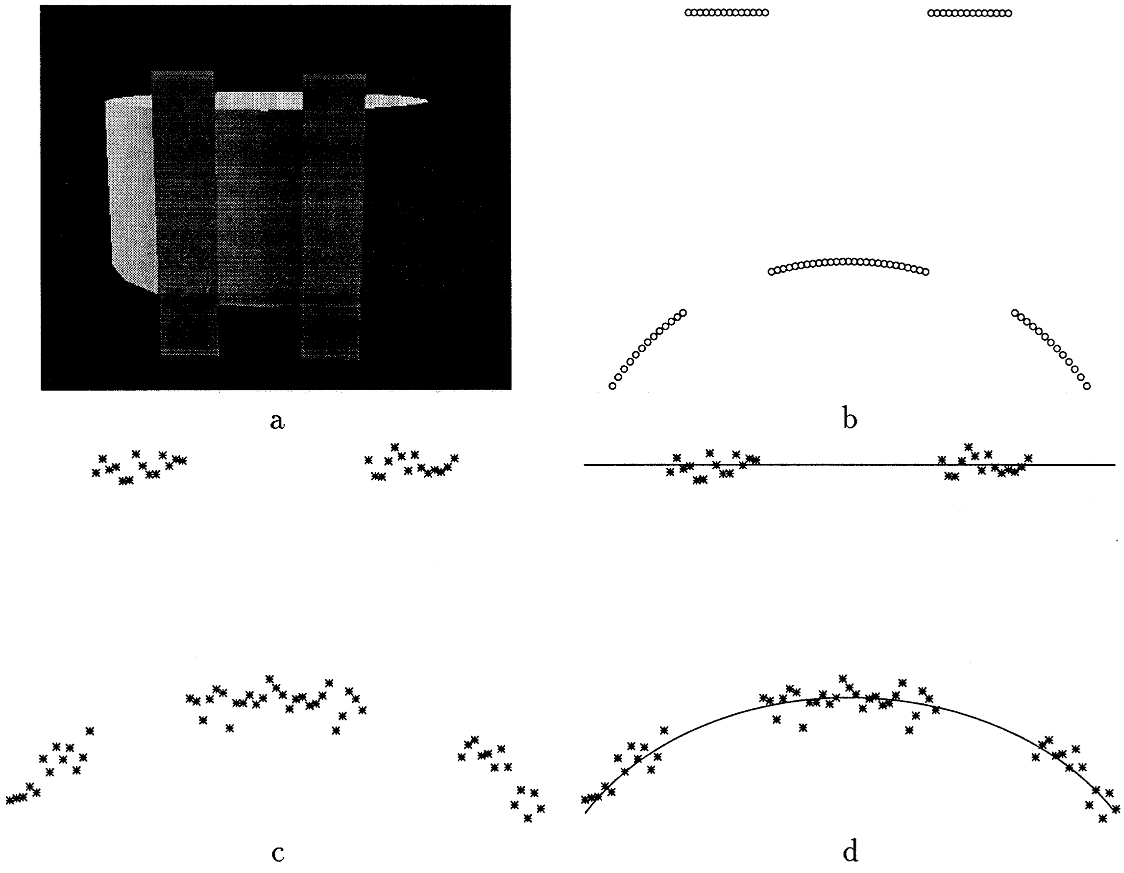


Figure 4-1: **a.** A simple three dimensional scene that can cause problems for existing motion segmentation algorithms. A cylinder is partially occluded by two bars. **b.** A cross section through the theoretical horizontal image velocity field caused by a moving camera. **c.** The same data as in (b) but with added Gaussian noise. In practice, the image velocity will be noisy. **d.** The desired description of the data.

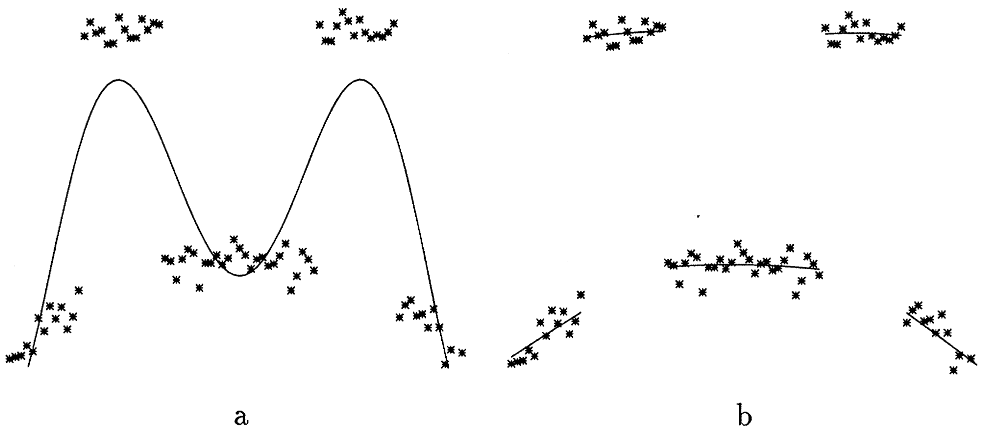


Figure 4-2: **a.** The fit of a single smooth curve to the data shown in figure 4-1c. Regularization causes heavy over-smoothing. **b.** Regularization with line processes. Fitting a smooth curve with discontinuities, or "line processes", causes two problems. First, there is no indication that the three occluded parts are part of a single object. Second, since no information is propagated between the occluded parts, the curvature of the cylinder is lost by the fit.

4.1 Introduction

Considerable progress in motion analysis has been achieved by systems that fit multiple global motion models to the image data (Darrell and Pentland, 1991; Jepson and Black, 1993; Irani and Peleg, 1992; Hsu et al., 1994; Ayer and Sawhney, 1995; Wang and Adelson, 1994). While differing in implementation, these algorithms share the goal of deriving from the image data a representation consisting of (1) a small number of global motion models and (2) a segmentation map that indicates which pixels are assigned to which model.

The advantages of these approaches over previous ones are twofold. First, by combining information over large region of the image, the local ambiguity of the image data is overcome and a reliable motion estimate can be found. Second, the derived segmentation map, in which individual pixels are grouped into perceptually salient parts, is useful for higher level processing (e.g. video database indexing, object recognition).

In order to segment images based on common motion, most existing algorithms assume that the motion of each model is described by a low dimensional parameterization. The two most popular choices are a six parameter affine model (Wang and Adel-

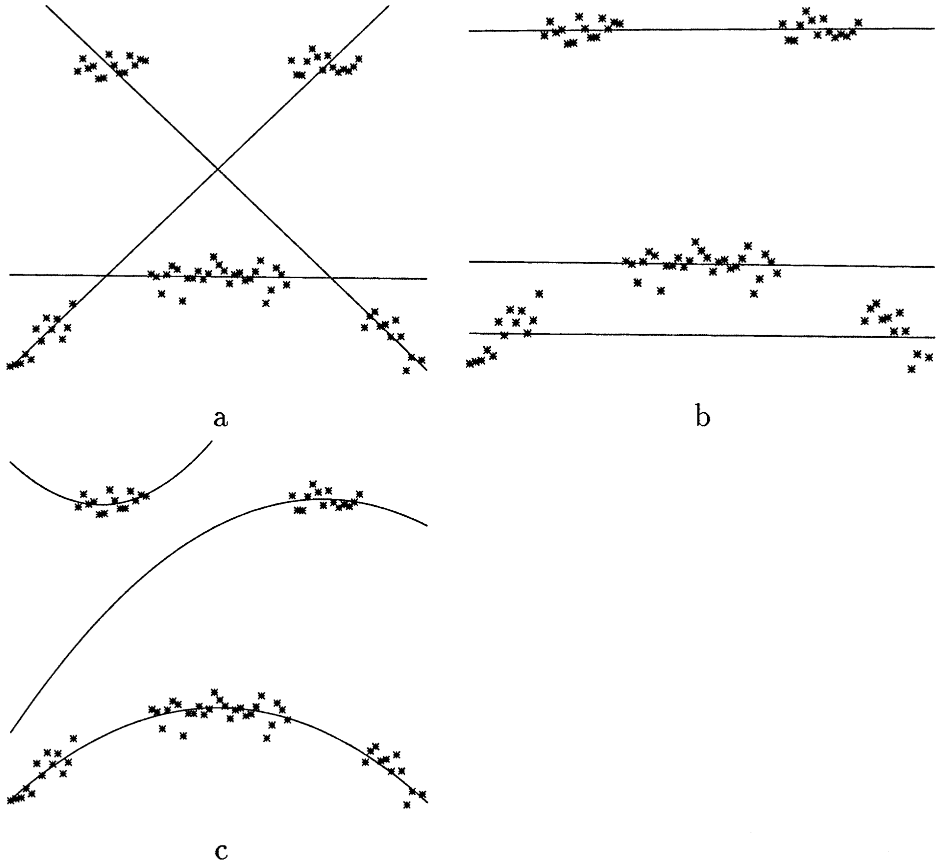


Figure 4-3: **a,b.** Two outputs of a multiple line fitting algorithm. Three lines are needed to achieve a reasonable fit, and the cylinder is broken apart. Various different solutions are found, and only two are shown. **c.** A result of extending the order of the models to quadratic. Although the model class is now rich enough to capture the data, estimation becomes unstable.

son, 1994; Weiss and Adelson, 1996) or an eight parameter projective model (Ayer and Sawhney, 1995; Irani and Peleg, 1992). Both of these parameterizations correspond to the rigid motion of a plane: the affine model assumes orthographic projection while the projective model assumes a perspective projection.

Despite the success of existing algorithms in segmenting image sequences, the assumption that motion segments correspond to rigid planar patches is obviously restrictive. Non-planar surfaces, or objects undergoing non-rigid motion cannot be grouped. In order for the motion segmentation map to be useful for higher level processing, these methods need to be extended so they can deal with non-planar surfaces and non-rigid motions.

Figure 4-1a shows a simple 3D scene that can cause problems for existing motion segmentation algorithms. A cylinder is partially occluded by two bars. Figure 4-1b shows a cross section of the horizontal component of optical flow when a camera, viewing the scene head on, is rotated horizontally about a distant point. The bars that are closer to the camera move fastest, and the velocity of points on the cylinder trace out a smooth curve. In practice, of course the velocity field will not be so perfect and Figure 4-1c shows the same data with added Gaussian noise.

One way to overcome the noisiness of the flow field is to use regularization to approximate the data with a single smooth function. Figure 4-2a shows the result of applying a typical regularization algorithm to the data shown in figure 4-1b. Although the noise is smoothed out, the fit suffers from heavy over-smoothing: the motions of the cylinder and the bars are averaged together. Figure 4-2b shows the output of a "regularization with discontinuities algorithm" (Terzopoulos, 1986) on the same data. Although this fixes the problem of oversmoothing, discontinuities are a bad model of occlusion (cf. (Marroquin, 1992; Darrell and Pentland, 1991; Madrasmi et al., 1993)): data in a scene containing multiple occluding objects is not generated by a single discontinuous function but rather multiple smooth functions interacting nonlinearly. The results of fitting a single discontinuous function causes two problems, which can be seen in the fit. First, there is no indication that the three parts of the cylinder are part of a single object. Second, because no information is propagated between the different fragments, the cylinder is fit with three nearly straight lines, rather than curved segments. The cylindrical shape is lost in the fit.

These limitations of regularization motivated much of the recent work in motion segmentation and led to the development of approaches that fit multiple global motion models to the data. How would parametric segmentation work on the data in figure 4-1? Figure 4-3 shows the output of a multiple parametric curve fitting algorithm to this data. The number of models was estimated automatically as in (Weiss and Adelson, 1996) (cf. (Wang and Adelson, 1994; Ayer and Sawhney, 1995)). When the curves are restricted to be lines, different outcomes are obtained depending on initial conditions, two of which are shown in figures 4-3a-b. Three lines are needed to achieve a reasonable fit, and the cylinder is fragmented.

What about using a quadratic model? In this case, the model class is rich enough to capture the data, but the estimation becomes unstable. Figure 4-3c shows a typical output. The instability of fitting higher order models causes each of the two bars to be fit with a parabola, an example of over-fitting. Although the correct fit is sometimes obtained, it is in no way favored over other erroneous interpretations.

The instability problems associated with increasing the dimensionality of parameterization are, of course, not limited to motion analysis or even to computer vision. It is generally accepted that one should avoid fitting high order polynomials to data. Multidimensional splines and regularization theory present an elegant alternative - the functions used in this approach are flexible enough to model the data yet avoid the instability associated with high order polynomials. Regularization theory has a long history of use in computer vision (Poggio et al., 1985) and has enjoyed considerable success, yet its disadvantages are well known. First, smoothness is simply a bad thing to assume over the whole image. Typically the image will contain multiple occluding objects, and assuming smoothness will lead to terrible estimates particularly in the regions of discontinuities. Second, calculating the regularized solution has typically involved highly iterative algorithms (e.g. (Geman and Geman, 1984; Terzopoulos, 1986; Marroquin, 1992; Madrasmi et al., 1993)) whose convergence may be excruciatingly slow.

Here we develop an approach to segmentation that is based on the assumption of *smoothness in layers*. Rather than assuming that the motion of the whole image varies smoothly, we assume that the motion of a given motion group or layer varies smoothly. We show how this leads to the notion of *nonparametric mixture estimation*, where the stability of the estimation process is ensured by means of a prior distribution on the class of flow fields. We present a variant of the EM algorithm that can perform the segmentation in a computationally feasible manner, and show how the algorithm is able to segment higher order flow fields while avoiding over-fitting.

4.2 Algorithm Description

4.2.1 Generative Model

The model assumes that the image data (the spatial and temporal derivatives of the sequence) were generated by K smooth motion groups. The velocity field of each group is drawn from a distribution where smooth velocities are more probable:

$$P(V) = \frac{1}{Z_1} e^{-\sum_{x,y} ||Dv(x,y)||} / \sigma_R^2 \tag{4.1}$$

Here Dv is a differential operator that penalizes fields that have strong derivatives:

$$Dv = \sum_{n=0}^{\infty} a_n \frac{\partial^n}{\partial x^n} v \tag{4.2}$$

We follow (Yuille and Grzywacz, 1989) in using $a_n = \sigma^{2n}/(n!2^n)$, although similar results are obtained with other choices.

The next stage is to generate a labeling of the image, i.e. a vector $L(x,y)$ at every location such that $L_k(x,y) = 1$ if and only if position x, y will be assigned to group k. The labelings are drawn from a Markov Random Field distribution:

$$P(L) = \frac{1}{Z_2} exp \left(\sum_{x,y,x',y'} w_{x,y,x',y'} L^t(x,y) L(x',y') \right) \tag{4.3}$$

The link weights $w_{x,y,x',y'}$ determine the distribution of labelings. For example setting $w_{x,y,x',y'} = 1$ for neighboring sites and zero otherwise makes labelings in which neighboring locations have similar labels more probable.

Now given the labeling and the velocity field of each group, the probability of observing $I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial t}$ at location (x, y) is given by:

$$P(I_x, I_t|L, V) = \exp(-\sum_k L_k(I_x^t v_k + I_t)^2/\sigma_N^2) \tag{4.4}$$

where, for clarity's sake, we have omitted the dependence of $L_k, I_x, I_y, I_t, v_k, L_k$ on (x, y) and σ_N is the expected level of noise in the sequence. Similar likelihood functions have been used for the single motion case in (Simoncelli et al., 1991; Luettgen et al., 1994). Note that here the likelihood depends on multiple velocities, but if $L_k(x, y)$ is known then the likelihood depends only on the velocity model to which a pixel is assigned.

4.2.2 Nonparametric mixture estimation

To estimate the parameters of this model we use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The algorithm iterates two steps: (1) the Expectation (or E) step in which the hidden labels are replaced with their conditional expectation and (2) the Maximization (or M) step in which the velocity fields are found by maximizing their posterior probability.

Previous implementations of the EM algorithm for motion segmentation are described in (Weiss and Adelson, 1996; Jepson and Black, 1993; Ayer and Sawhney, 1995). Two aspects of the algorithm used here are similar to the implementation described in (Weiss and Adelson, 1996) and will only be described briefly:

- The number of models is estimated automatically, by initializing the algorithm with more models than will be needed. The algorithm merges redundant models and the final number of models found depends on the parameter σ_N.

- The MRF priors on the labelings make an exact E step computationally expensive and hence a consistent approximation to the MRF distribution is used for which an exact E step can be computed efficiently.

4.2.3 Estimating smooth flow fields using Green's functions

The distinguishing feature of our algorithm in comparison to previous EM based approaches is in the M step. It requires finding, for each model, the dense flow field that maximizes the conditional posterior probability, or equivalently minimizes:

$$\begin{aligned} J_k(V) &= \sum_{x,y} \hat{L}_k(x,y) \left(I_x^t(x,y)v(x,y) + I_t(x,y)\right)^2 \\ &\quad +\lambda \sum_{x,y} \|Dv(x,y)\| \end{aligned} \tag{4.5}$$

where the parameter λ is determined by the ratios of σ_N and σ_R in the generative model. $\hat{L}_k(x,y)$ is the "filled in" estimate for the labeling at location (x,y). It is these weights in equation 4.5 that cause the estimated dense flow to differ from model to model.

Since the nonparametric EM algorithm calls for minimizing equation 4.5 at every iteration for all models, this approach can only be computationally feasible if the minimization can be performed efficiently. We now show how this can be done.

Using the method of Green's Functions (cf. (Yuille and Grzywacz, 1989; Girosi et al., 1993)) it can be shown that the optimal velocity field V_k^* is a linear combination of basis flow fields,$B_i(x,y)$:

$$V_k^*(x,y) = \sum_i \alpha_i B_i(x,y) \tag{4.6}$$

There is a basis flow field centered at every pixel where the image gradient is nonzero:

$$B_i(x,y) = G(x-x_i, y-y_i)\left[I_x(x_i,y_i), I_y(x_i,y_i)\right]^t \tag{4.7}$$

The scalar valued function $G(x,y)$ is the Green's function corresponding to the differential operator D in equation 4.5 (cf. (Strang, 1986)). It is a solution to:

$$D^* DG = \delta(x, y) \tag{4.8}$$

For the differential operator used here, the Green's function is a two dimensional Gaussian (Yuille and Grzywacz, 1989). The coefficients α are the solution to the linear system:

$$(WM + \lambda I)\alpha = WY \tag{4.9}$$

With M_{ij} is given by the scalar product of the basis field centered on pixel i and the gradient at pixel j, Y_i is simply the temporal derivative at pixel i and W is a diagonal matrix whose diagonal elements determine the weight of a pixel in estimating model parameters $W_{ii} = \hat{L}_k(x_i, y_i)$.

Although equation 4.9 gives a closed form solution for the optimal velocity field for each model, its solution is still computationally prohibitive as it requires solving a linear system whose rank is equal to the number of nonzero gradients in the image. Are we then back to square one? No, because a remarkably good suboptimal solution can be found using this method in a computationally feasible way.

The suboptimal solution is obtained by using only a subset of the basis fields in equation 4.6. Denote by N the number of basis fields in the reduced expansion, then the N coefficients are a solution to:

$$(M^t WM + \lambda R)\alpha = M^t WY \tag{4.10}$$

where M_{ij} is again given by the scalar product of the basis field centered on pixel i and the gradient at pixel j, R is a NxN submatrix of M in which only the pixels which have basis functions centered on them are used, and W and Y are as before. Note that equation 4.10 is of rank N independent of the number of pixels. Note also the term λR in the left hand side of equation 4.10. It is this term that imposes the prior distribution and makes the estimation well posed regardless of the dimensionality of the parameter vector α. In general, the solution obtained by solving equation 4.10

will be different from one obtained by simply assuming the flow field is parameterized by a spline basis set (e.g. (Szeliski and Shum, 1995)). Finally, note that the reduced rank of the system is obtained by using only a subset of the basis fields, *not* by using a subset of the gradient constraints. The solution of the system gives the flow field spanned by the reduced basis set that best satisfies the gradient constraints at all pixels.

The difference between the optimal and the suboptimal solution depends on the image data, the differential operator D and the choice of subsets. In practice, we have found the difference to be negligible when 50 basis fields are used, chosen so that they are equally spaced on the image. To get an intuition regarding the optimal and suboptimal solutions, we generated a synthetic sequence by warping the image shown in figure 4-4a according to the superimposed flow. Figure 4-4b shows cross sections from the estimated velocity fields. The suboptimal solution is plotted with crosses, and the optimal one is plotted with circles. The solutions are indistinguishable. On a R4400 Silicon Graphics workstation, solving equation 4.10 to calculate the suboptimal solution took less than 1/100 of a second, while solving equation 4.9 to calculate the optimal solution took over an hour.

Although we have used here the differential operator suggested in (Yuille and Grzywacz, 1989) the exact same method can be used with other differential operators. For example, we have been able to solve the Horn and Schunck (Horn and Schunck, 1981) equations in closed form using this method.

4.2.4 Algorithm summary

To summarize, the statistical assumptions about the generative model are characterized by four numbers: σ, λ which embody the smoothness assumption, σ_N the assumed level of noise in the sequence and $w_{xyx'y'}$ which specifies the probability that a pixel will belong to a different model than its four neighbors.

Given these assumptions and spatiotemporal derivatives computed over the image, we use a computationally efficient EM algorithm to calculate number of models, the segmentation of the image and a smooth dense flow field for every model.

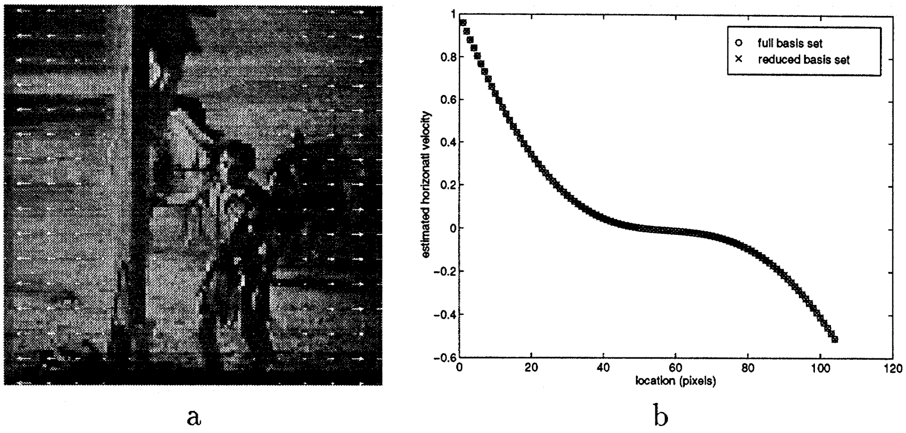


Figure 4-4: Using the method of Green's functions, a closed form solution can be found for fitting a smooth dense flow field to the image data. A suboptimal solution, which is computationally efficient can also be found. We have found the difference between the optimal and suboptimal solutions to be negligible. **a.** A frame from a test sequence. A second frame was generated by warping this frame with the superimposed flow field. **b.** Cross sections from the estimated velocity field using the full basis function set (circles) and using only 50 basis functions (crosses). The two solutions are indistinguishable.

4.3 Results

Before showing the results of our motion segmentation algorithm, we show the performance of a similar 1D nonparametric mixture estimation algorithm on the data discussed in the introduction and shown in figure 4-1c. Although some of the problems characteristic of motion segmentation are not present in 1D (e.g. the aperture problem), we choose to first illustrate the performance on a 1D problem because it enables us to display the evolution of the model's estimates. Figure 4-5 shows the line fits and estimated labelings, $\hat{L}_k(x)$, as a function of iteration. Note that although the label $L_k(x)$ are assumed to be binary, their "filled in" estimates $\hat{L}_k(x)$ are continuous valued and lie between zero and one. The algorithm is initialized with four curves each of which has is initially assigned a random subset of the data. Hence the initial fits are nearly identical. After six iterations, when the algorithm converges, two of the models are merged and only two unique models are needed to explain the data.

Compare the fit obtained by our algorithm to those discussed in the introduction. Unlike the regularization with discontinuities fit in figure 4-2b, our algorithm combines

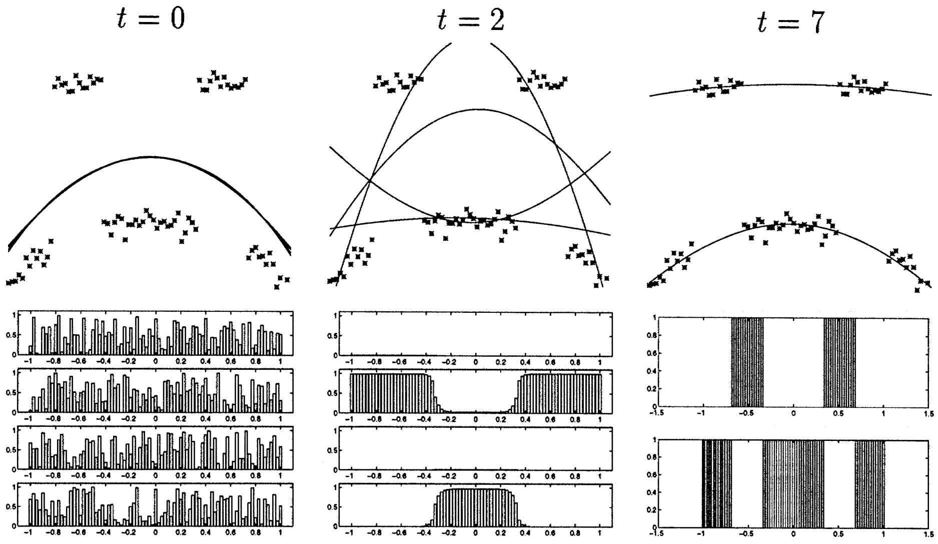


Figure 4-5: The performance of our algorithm on the data shown in 4-1c. The model is randomly initialized with four curves and automatically decides that two curves are sufficient in this case. The algorithm converges in seven iterations. Note that the curvature of the cylinder is correctly estimated.

information across the different portions of the occluded cylinder and the curvature of the cylinder is apparent in the fit. Since each of the models is flexible enough, our algorithm can achieve a good fit with just two curves, unlike the line fit shown in figure 4-3. Since our algorithm uses a prior favoring smooth fits, it does not over-fit as does the quadratic fit shown in figure 4-3c.

We now show an example of the full 2D motion segmentation algorithm. We generated a synthetic image sequence modeled after the scene in figure 4-1a. Figure 4-6a shows a single frame from the sequence. A textured cylinder is partially occluded by two textured bars, and the camera is rotating about a distant center and translating. The camera was assumed to be orthographic and the translation was such that the mean horizontal velocity of the image was zero. Similar to the 1D case discussed earlier, this sequence is hard to segment using parametric approaches. Figures 4-6b-c show the output of our algorithm – it correctly estimates the number of models and the segmentation. The high quality velocity field obtained using our method enables us to reconstruct a three dimensional surface for each segment output (assuming orthography, this is simply the horizontal component of the derived dense

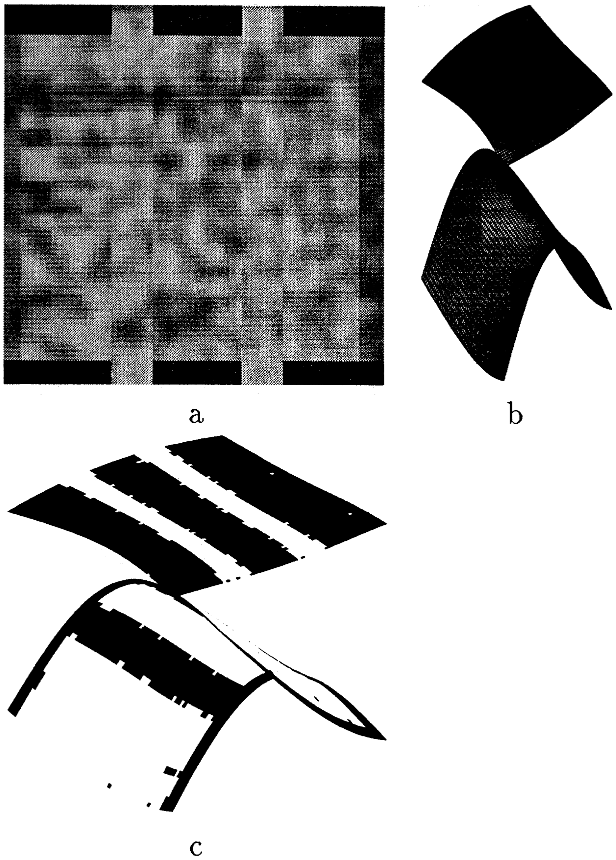


Figure 4-6: **a.** A single frame from the cylinder sequence. A textured cylinder is partially occluded by two textured bars, and the camera is rotating about a distant center. **b.** Reconstructed three dimensional surfaces obtained from the horizontal dense velocity fields estimated by our algorithm. **c.** The segmentation maps displayed on top of the surfaces, indicating the opacity of each layer.

flow fields). Figure 4-6b shows the surfaces obtained in this way, and Figure 4-6c shows the segmentation maps displayed on top of the surfaces, indicating the opacity of each layer.

We have found that sequences which are easily segmented using parametric motion models are also segmented using our approach. This is not surprising – the low dimensional motion models are often smooth and hence favored as segmentations by our model. Figure 4-7a shows a single frame from a sequence that was segmented using translational models in (Jepson and Black, 1993; Darrell and Pentland, 1991). A person is moving behind a plant. Figure 4-7b shows the segmentation derived by our algorithm. The parameter settings are identical to those used in the cylinder

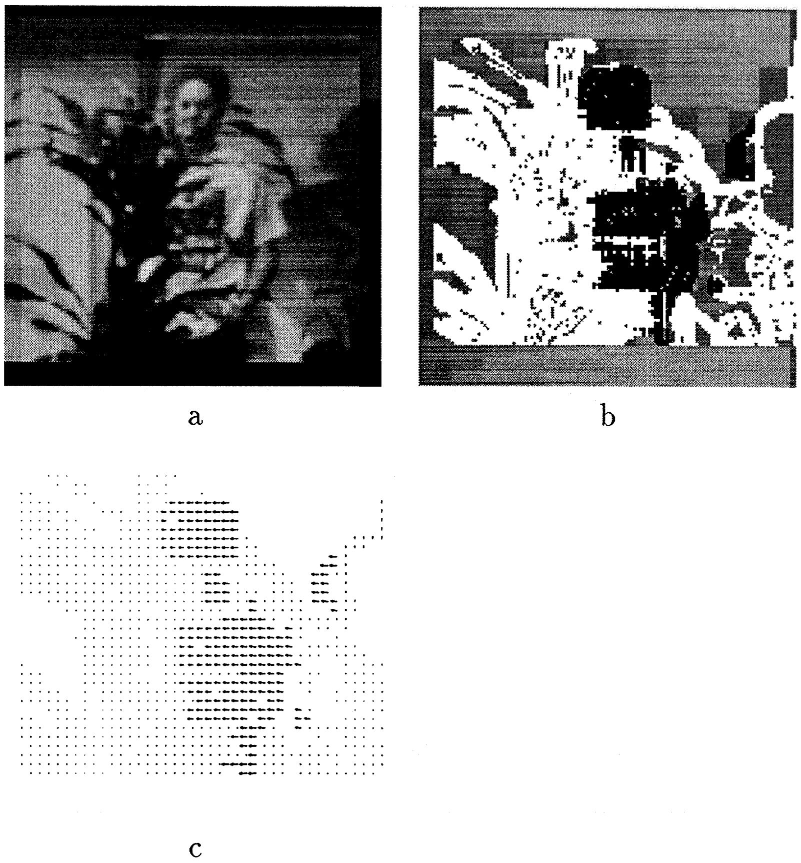


Figure 4-7: **a.** The plant sequence. A person is moving behind a plant. **b.** The segmentation found by our algorithm. Pixels belonging to the person are grouped together. **c.** The velocity estimate obtained by plotting at each pixel the velocity of the model to which that pixel is assigned.

sequence. The number of models is correctly estimated and the different fragments corresponding to the person are grouped together. Figure 4-7c shows the estimated velocities.

Figure 4-8a shows a single frame from the MPEG flower-garden sequence that was segmented using planar models in (Wang and Adelson, 1994; Ayer and Sawhney, 1995; Weiss and Adelson, 1996). Since this sequence contains large motions, we replaced the temporal derivative in equation 4.4 with a calculated normal velocity at each pixel. The normal velocity was calculated using a coarse-to-fine method (cf. (Bergen et al., 1992)). The other aspects of the algorithm were identical to those used in previous sequences.

For the value of σ_N used two segments are found, one corresponding to the tree (shown in figure 4-8b) and the other corresponding to the rest of the image. An advantage of using nonparametric models for the segmentation, is that the nonplanarity of the scene can be captured in the output. Figure 4-8c shows a cross section through the horizontal flow recovered by our algorithm (since the camera motion is roughly horizontal, this flow is approximately related to distance from the camera). The cross section is taken at the position indicated by the dotted line in figure 4-8a. Note that the motions of the flower bed and the tree are smooth, curved functions. This type of structure can be easily captured in a nonparametric technique, but is lost when segments are assumed to be 3D planes. As a result, when we use the estimated motions to align the two frames, we obtain a noticeably better alignment with the nonparametric segmentation technique as compared to affine segmentation.

4.4 Discussion

Motion segmentation algorithms are often categorized as "direct" or "indirect" based on whether they fit models directly to the image data (e.g. (Ayer and Sawhney, 1995)) or to local optical flow measurements (e.g. (Wang and Adelson, 1994)). The particular implementation presented here would be classified as direct, since the models are fit to the spatiotemporal derivatives (see equation 4.4). However, the framework we have developed here is in no way restricted to spatiotemporal derivatives and can also be applied to local optical flow measurements, in cases when an indirect method is judged to be advantageous.

Our generative model assumes that for every pixel, there exists a motion model that generated the spatiotemporal derivatives at that pixel. This formulation ignores the accretion and deletion of pixels at occlusion boundaries that give rise to spatiotemporal data that is not well explained by any of the motion models. In current work, we are exploring the use of outlier models to deal with those pixels (cf. (Jepson and Black, 1993)).

The preceding discussions highlight the relationship between our approach and existing segmentation algorithms. Our approach fits a dense smooth flow field for

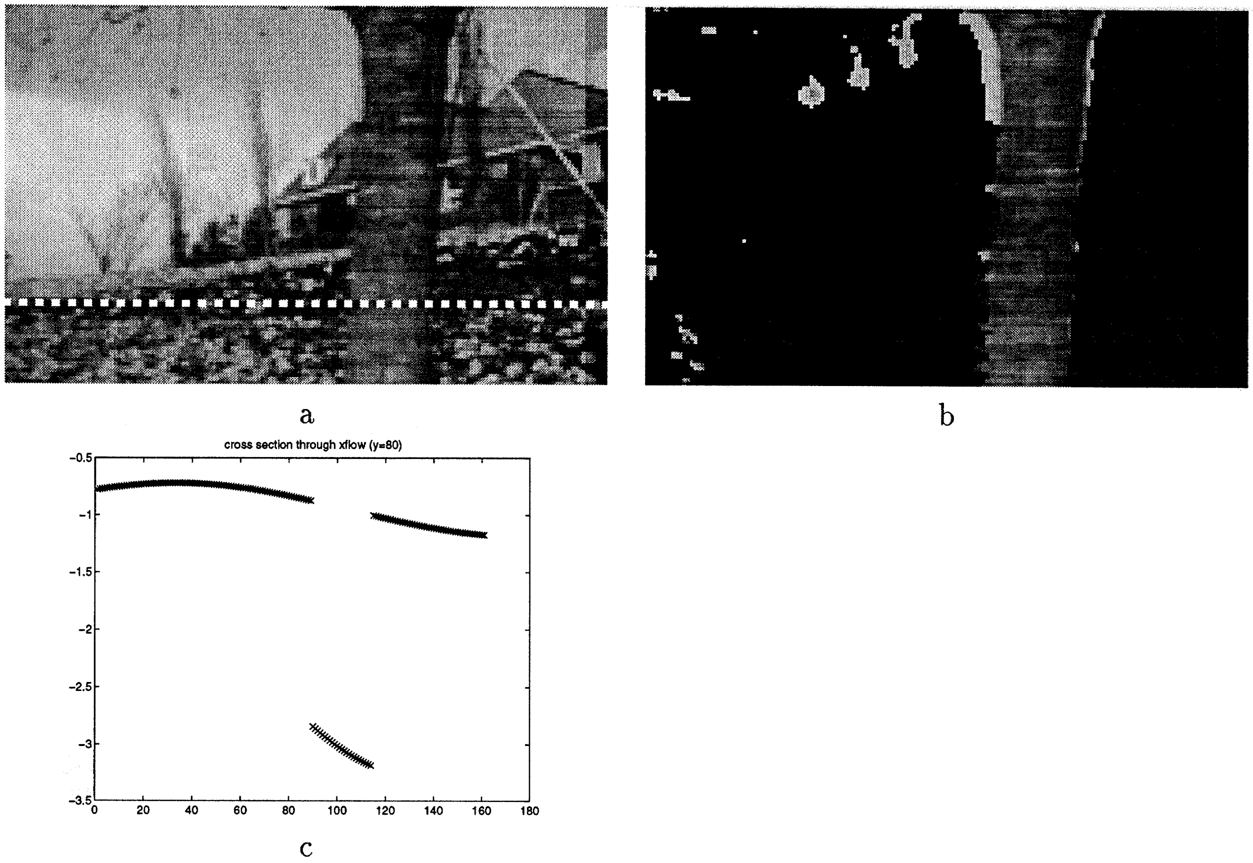


Figure 4-8: **a.** The flower garden sequence. The camera is translating approximately horizontally. **b.** The segmentation found by our algorithm. Two segments are found - one corresponding to the tree (shown) and another corresponding to the rest of the image. **c.** A cross section through the horizontal flow field taken at the dotted line in a. Note that the algorithm correctly finds the nonplanar motions of the flower bed and the tree.

every segment, and this allows us to segment non-planar surfaces or objects undergoing non-rigid motions. However, our approach shares the basic structure of existing parametric segmentation algorithms, and thus when dealing with questions of model selection, large motions and outlier rejection, we can build on the progress made by existing algorithms.

The distinction between parametric and nonparametric estimation may seem rather arbitrary. Indeed, the dense flow field by which we represent the motion of each group may be thought of as a parametric description with the number of parameters equal to the number of pixels. However, there is a fundamental difference between the two approaches. The difference is not in the number of free parameters but rather lies in what is responsible for making the estimation well posed. In parametric approaches, this is accomplished by assuming a *small number* of unknowns, while in nonparametric approaches the well-posedness is a result of assuming a *prior distribution* over the unknowns. In this work, we assumed a prior distribution where the probability of a flow field is inversely related to its smoothness and showed how to efficiently maximize the posterior probability under this assumption. An advantage of the nonparametric mixture framework developed here, is that other types of prior distributions can be easily incorporated in place of the smoothness assumption. Thus this framework can be used to investigate what assumptions are necessary to achieve stable segmentation of arbitrary image sequences.

4.5 Conclusion

Existing motion segmentation algorithms are able to segment image sequences by restricting the motion of each segment to lie in a low dimensional subspace. This approach has inherent limitations. If the subspace is small then it is too restrictive and cannot group together pixels undergoing more complex motions. If the subspace is rich enough to capture complex motions, the dimensionality is large and the estimation becomes unstable.

Existing regularization approaches avoid some of the shortcomings of parametric models but introduce new problems. The assumption of smoothness over the whole image leads to erroneous estimates in any scene containing multiple objects, and the solution involves slow, iterative calculations. The addition of "line processes" to the regularization framework only partially addresses these problems: line processes are a bad model for occlusion, thus disabling the propagation of information between occluded fragments, and the computational cost associated with these algorithms is even more prohibitive.

Here we have developed a new approach that builds on the recent progress made in statistically based segmentation. We have presented a generative model that embodies a prior towards smoothness, but smoothness in a layer and not smoothness over the whole image. We have shown how this leads to nonparametric mixture estimation and developed a variant of the EM algorithm that can efficiently perform segmentation under this assumption. By deriving a closed form solution to the smooth motion problem, we are able to avoid the slow iterative calculations of traditional approaches. Based on the successful performance of our algorithm on synthetic and real image sequences, we are optimistic that this framework will also be useful for other segmentation tasks in computational vision.

Chapter 5

Conclusions

The main goal of this thesis was to understand how the human visual system solves the "integration versus segmentation" dilemma. Using the method of computational modeling we have shown that a large number of percepts are predicted by a Bayesian estimator that incorporates a small number of assumptions. We have also presented a computer vision segmentation algorithm that is based on similar assumptions and illustrated its performance on real and synthetic image sequences.

The ability of our Bayesian model to account for such a wide range of phenomena suggests the following conclusions regarding motion perception in the human visual system:

- The system assumes image sequences may be noisy or ambiguous. The initial stages of motion processing do not merely extract local velocity estimates but also indicate the degree to which these estimates are ambiguous.

- The system assumes that velocity fields tend to be slow and smooth. The motion percept is based on combining this prior assumption with the local estimates.

- The combination takes into account the uncertainty of the local estimates. Estimates with high uncertainty (e.g. based on low contrast, or a single orientation) are given low weight in comparison to other estimates and in comparison to the prior assumption.

- The assumption of slow and smooth velocity fields does not imply that velocity changes smoothly over the entire scene. Rather the system assumes a small number of layers in the scene, and assumes that the velocity field of each layer is slow and smooth.

- The decision of whether to integrate or segment motion measurements is not based solely on motion information. Static segmentation cues play an important role in this decision.

Throughout this thesis we did not address the question of how these assumptions and constraints may be mapped on to what is known about the functional architecture of the visual system. This appears to be a promising area for future research. For example: how can the uncertainty of an estimate by represented in early visual areas? What biological mechanism is capable of combining together uncertain estimates according to their uncertainty? How can something like the assignment of locations to surfaces be represented in the brain ?

Obviously, the ability of computational modeling to answer these questions is inherently limited. An integrative approach that combines psychophysics, neurophysiology and modeling is required. We would like to reemphasize, however, the importance of a computational theory over a verbal theory in this integrative approach. As we discuss in the thesis, the suggestion that humans may have a preference for slow velocities goes back to the beginning of the century. But in the absence of a computational formalization, it is far from obvious that this assumption would lead to percepts such as the movement of a rhombus in the vector average direction, the apparent nonrigidity of fat ellipses and the incoherence of plaids with narrow angles. The formal computational model bridges the gap between abstract assumptions on the one hand, and a predicted percept for a gray level image sequence on the other hand. We hope that this thesis will serve as motivation for continuing to pursue this method for understanding perception.

Bibliography

Adelson, E. and Movshon, J. (1982). Phenomenal coherence of moving visual patterns. *Nature*, 300:523–525.

Adelson, E. H. and Bergen, J. R. (1986). The extraction of spatio-temporal energy in human and machine vision. In *Proceedings of the Workshop on Motion: Representation and Analysis*, pages 151–155, Charleston, SC.

Adelson, E. H. and Movshon, J. A. (1983). The perception of coherent motion in two-dimensional patterns. In *ACM Siggraph and Sigart interdisciplinary workshop on Motion: Representation and Perception*, pages 11–16, Toronto.

Alais, D., Wenderoth, P., and Burke, D. (1994). The contribution of one-dimensional motion mechanisms to the perceived direction of drifting plaids and their aftereffects. *Vision Research*, 34(14):1823–1834.

Anstis, S. (1990). Imperceptible intersections: the chopstick illusion. In Blake, A. and Troscianko, T., editors, *AI and the eye*. John Wiley.

Ayer, S. and Sawhney, H. S. (1995). Layered representation of motion video using robust maximum likelihood estimation of mixture models and MDL encoding. In *Proc. Int'l Conf. Comput. Vision*, pages 777–784.

Bergen, J., Anandan, P., Hana, K., and al, R. H. (1992). Hierarchial model-based motion estimation. In *Proc. Second European Conf. on Comput. Vision*, pages 237–252, Santa Margherita Ligure, Italy.

Black, M. J. and Rangarajan, A. (1994). The outlier process: Unifying line processes and robust statistics. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 15–22, Seattle, Washington.

Bowns, L. (1996). Evidence for a feature tracking explanation of why type II plaids move in the vector sum directions at short directions. *Vision Research*, 36(22):3685–3694.

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in Neuroscience*, 16:263–268.

Bressan, P., Ganis, G., and Vallortigara, G. (1993). The role of depth stratification in the solution of the aperture problem. *Perception*, 22:215–228.

Broomhead, D. and Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.

Bulthoff, H., Little, J., and Poggio, T. (1989). A parallel algorithm for real-time computation of optical flow. *Nature*, 337(6207):549–553.

Burke, D. and Wenderoth, P. (1993). The effect of interactions between one-dimensional component gratings on two dimensional motion perception. *Vision Research*, 33(3):343–350.

Burt, P. and Sperling, G. (1981). Time, distance, and feature trade-offs in visual apparent motion. *Psychological Review*, 88(2):171–195.

Darrell, T. and Pentland, A. (1991). Robust estimation of a multi-layered motion representation. In *Proc. IEEE Workshop on Visual Motion*, pages 173–178, Princeton, New Jersey.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38.

Durbin, R., Szeliski, R., and Yuille, A. (1989). An analysis of the elastic net approach to the travelling salesman problem. *Neural Computation*, 1(3):348–358.

Farid, H. and Simoncelli, E. P. (1994). The perception of coherence in square-wave plaids. *Investigative Opthamology and Visual Science*, 35.

Farid, H., Simoncelli, E. P., Bravo, M., and Schrater, P. (1995). Effects of contrast and period on perceived coherence of moving square-wave plaids. *Investigative Opthamology and Visual Science*, 36(4).

Fennema, C. and Thompson, W. (1979). Velocity determination in scenes containing several moving objects. *Computer Graphics and Image Processing*, 9:301–315.

Ferrera, V. and Wilson, H. (1990). Perceived direction of moving two-dimensional patterns. *Vision Research*, 30:273–287.

Freeman, W. T. (1992). *Steerable Filters and Local Analysis of Image Structure.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. PAMI*, 6(6):721–741.

Girosi, F., Jones, M., and Poggio, T. (1993). Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. AI Memo No: 1430, MIT AI Lab.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269.

Girosi, F., Poggio, T., and Caprile, B. (1990). Extensions of a theory of networks for approximation and learning: outliers and negative examples. AI Memo No: 1220, MIT AI Lab.

Grzywacz, N. and Yuille, A. (1991). Theories for the visual perception of local velocity and coherent motion. In Landy, J. and Movshon, J., editors, *Computational models of visual processing.* MIT Press, Cambridge, Massachusetts.

Heeger, D. J. and Simoncelli, E. P. (1991). Model of visual motion sensing. In Harris, L. and Jenkin, M., editors, *Spatial Vision in Humans and Robots.* Cambridge University Press.

Hildreth, E. C. (1983). *The Measurement of Visual Motion.* MIT Press.

Horn, B. K. P. (1986). *Robot Vision.* The MIT Press, Cambridge, MA.

Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artif. Intell.*, 17(1–3):185–203.

Hsu, S., Anandan, P., and Peleg, S. (1994). Accurate computation of optical flow by using layered motion representation. In *Proc. 12th Int'l Conf. Pattern Recog.*

Hutchinson, J., Koch, C., Luo, J., and Mead, C. (1988). Computing motion using analog and binary resistive networks. *IEEE Computer magazine*, 21:52–64.

Irani, M. and Peleg, S. (1992). Image sequence enhancement using multiple motions analysis. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 216–221, Champaign, Illinois.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Jasinschi, R., Rosenfeld, A., and Sumi, K. (1992). Perceptual motion transparency: the role of gemotrical information. *Journal of the Optical Society of America A*, 9(11):1865–1879.

Jepson, A. and Black, M. J. (1993). Mixture models for optical flow computation. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 760–761, New York.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.

Kim, J. and Wilson, H. (1993). Dependence of plaid motion coherence on component grating directions. *Vision Research*, 33(17):2479–2489.

Knill, D. and Richards, W. (1996). *Perception as Bayesian Inference.* Cambridge University Press.

Lorenceau, J. and Shiffrar, M. (1992). The influence of terminators on motion integration across space. *Vision Research*, 32:263–273.

Lorenceau, J., Shiffrar, M., Wells, N., and Castet, E. (1992). Different motion sensitive units are involved in recovering the direction of moving lines. *Vision Research*, 33(9):1207–1217.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Image Understanding Workshop*, pages 121–130.

Luettgen, M. R., Karl, W. C., and Willsky, A. S. (1994). Efficient multiscale regularization with application to the computation of optical flow. *IEEE Transactions on image processing*, 3(1):41–64.

Madrasmi, S., Kersten, D., and Pong, T. (1993). Multi-layer surface segmentation using energy minimzation. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 774–775.

Marr, D. (1982). *Vision.* H. Freeman and Co.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neuroscience Research Progress Bulletin*, 15:470–488.

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.

Marroquin, J., Mitter, S., and Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82:76–89.

Marroquin, J. L. (1992). Random measure fields and the integration of visual information. *IEEE Transactions on Systems, Man and Cybernetics*, 22(4):705–716.

Mingolla, E., Todd, J., and Norman, J. (1992). The perception of globally coherent motion. *Vision Research*, 32(6):1015–1031.

Movshon, A., Adelson, E., Gizzi, M., and Newsome, W. (1986). The analysis of moving visual patterns. *Experimental Brain Research*, 11:117–152.

Musatti, C. (1924). Sui fenomeni stereocinetici. *Archivio Italiano di Psicologia*, 3:105–120.

Musatti, C. (1975). Stereokinetic phenomena and their interpretation. In Darcais, G. B. F., editor, *Studies in Perception: Festschrift for Fabio Metelli.* Martello - Giunti, Milano.

Nakayama, K. and Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, 257:1357–1363.

Nakayama, K. and Silverman, G. H. (1988a). The aperture problem - I: Perception of nonrigidity and motion direction in translating sinusoidal lines. *Vision Research*, 28:739–746.

Nakayama, K. and Silverman, G. H. (1988b). The aperture problem - II: Spatial integration of velocity information along contours. *Vision Research*, 28:747–753.

Nowlan, S. J. and Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *The Journal of Neuroscience*, 15(2):1195–1214.

Poggio, T. and Girosi, F. (1989). A theory of networks for approximation and learning. AI Memo No: 1140, MIT AI Lab.

Poggio, T. and Reichardt, W. (1973). Considerations on models of movement detection. *Kybernetik*, 13:223–227.

Poggio, T., Torre, V., and Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317:314–319.

Qiyan, N., Andersen, R., and Adelson, E. (1994). Transparent motion perception as detection of unbalanced motion signals. I psychophysics. *The journal of neuroscience*, 14(12):7357–7366.

Ramachandran, V. (1990). Visual perception in people and machines. In Blake, A. and Troscianko, T., editors, *AI and the eye*. John Wiley.

Reichardt, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In Rosenblith, W. A., editor, *Sensory Communication*. Wiley.

Rodman, H. and Albright, T. (1989). Single-unit analysis of pattern motion selective properties in the middle temporal visual area MT. *Experimental Brain Research*, 75:53–64.

Rose, K., Gurewitz, F., and Fox, G. (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65:945–948.

Rubin, N. and Hochstein, S. (1993). Isolating the effect of one-dimensional motion signals on the perceived direction of moving two-dimensional objects. *Vision Research*, 33:1385–1396.

Sajda, P. and Finkel, L. H. (1994). Intermediate-level visual representations and the construction of surface perception. *Journal of Cognitive Neuroscience.*

Shiffrar, M., Xiaojun, L., and Lorenceau, J. (1995). Motion integration across differing image features. *Vision Research*, 35(15):2137–2146.

Shimojo, S., Silverman, G., and Nakayama, K. (1989). Occlusion and the solution to the aperture problem for motion. *Vision Research*, 29:619–626.

Shizawa, M. (1993). Multi-valued standard regularization theory (1): Global reconstruction of multiple transparent surfaces via massively parallel relaxation algorithms. Technical Report TR-H-037, ATR Human Information Processing Laboratories.

Shizawa, M. and Mase, K. (1991). A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 289–295, Maui, Hawaii.

Simoncelli, E., Adelson, E., and Heeger, D. (1991). Probability distributions of optical flow. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 310–315.

Simoncelli, E. and Heeger, D. (1992). A computational model for perception of two-dimensional pattern velocities. *Investigative Opthamology and Vision Research*, 33.

Simoncelli, E. and Heeger, D. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.

Simoncelli, E. P. (1993). *Distributed Representation and Analysis of Visual Motion.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts of Technology, Cambridge.

Stone, L., Watson, A., and Mulligan, J. (1990). Effect of contrast on the perceived direction of a moving plaid. *Vision Research*, 30(7):1049–1067.

Stoner, G., Albright, T., and Ramachandran, V. (1990). Transparency and coherence in human motion perception. *Nature*, 344:153–155.

Strang, G. (1986). *Introduction to Applied Mathematics.* Wellesley-Cambridge.

Szeliski, R. and Shum, H.-Y. (1995). Motion estimation with quadtree splines. In *Proc. Int'l Conf. Comput. Vision*, pages 757–762.

Tenenbaum, J. B. and Todorov, E. V. (1995). Factorial learning by clustering features. In Tesauro, G., Touretzky, D., and Leen, K., editors, *Advances in Neural Information Processing Systems 7.*

Terzopoulos, D. (1986). Regularization of inverse visual problems involving discontinuities. *IEEE Trans. PAMI*, 8:413–424.

Thompson, P., Stone, L., and Swash, S. (1996). Speed estimates from grating patches are not contrast normalized. *Vision Research*, 36(5):667–674.

Tikhonov, A. and Arsenin, V. (1977). *Solution of Ill-Posed problems.* W.H. Winston, Washington DC.

Torr, P. (1998). Geometric motion segmentation and model selection. *Phil. Trans. R. Soc. Lond. A.*

Ullman, S. (1979). *The interpretation of visual motion.* The MIT Press.

Vallortigara, G. and Bressan, P. (1991). Occlusion and the perception of coherent motion. *Vision Research*, 31(11):1967–1978.

Wallach, H. (1935). Ueber visuell whargenommene bewegungrichtung. *Psychologische Forschung*, 20:325–380.

Wallach, H., Weisz, A., and Adams, P. A. (1956). Circles and derived figures in rotation. *American Journal of Psychology*, 69:48–59.

Wang, J. Y. A. and Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, 3(5):625–638.

Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, pages 520–527.

Weiss, Y. (1998). Phase transitions and perceptual organization of video sequences. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems 10.*

Weiss, Y. and Adelson, E. (1995). Adventures with gelatinous ellipses. *Perception*, 24(supplement).

Weiss, Y. and Adelson, E. H. (1996). A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, pages 321–326.

Weiss, Y. and Adelson, E. H. (1998). Slow and smooth: a Bayesian theory for the combination of local motion signals in human vision. Technical Report 1624, MIT AI lab.

Welch, L. (1989). The perception of moving plaids reveals two processing stages. *Nature*, 337:734–736.

Wilson, H., Ferrera, V., and Yo, C. (1992). A psychophysically motivated model for two-dimensional motion perception. *Visual Neuroscience*, 9:79–97.

Wilson, H. and Kim, J. (1994). A model for motion coherence and transparency. *Visual Neuroscience*, 11:1205–1220.

Wuerger, S., Shapley, R., and Rubin, N. (1996). On the visually perceived direction of motion by hans wallach: 60 years later. *Perception*, 25:1317–1367.

Yo, C. and Wilson, H. (1992). Perceived direction of moving two-dimensional patterns depends on duration, contrast, and eccentricity. *Vision Research*, 32(1):135–147.

Yuille, A. L. and Grzywacz, N. M. (1989). a mathematical analysis of the motion coherence theory. *Int'l J. Comput. Vision*, 3:155–175.

Zhang, J., Modestino, W., and Langan, D. (1994). Maximum-likelihood parameter estimation for unsupervised model-based image segmentation. *IEEE Transactions on Image Processing*, 3(4):404–420.