Computing Spatiotemporal Relations for Dynamic Perceptual Organization†

Mark Allmen

Sandia National Laboratories
P.O. Box 969 ORG 8351
Livermore, CA 94511

Phone: (510) 294-1248
FAX: (510) 294-1004

Charles R. Dyer

Computer Sciences Department
University of Wisconsin
Madison, Wisconsin 53706

† The support of the National Science Foundation under Grant Number IRI-9022608 is gratefully acknowledged.

Dynamic Perceptual Organization

Abstract

To date, the overwhelming use of motion in computational vision has been to recover the three-dimensional structure of the scene. We propose that there are other, more powerful, uses for motion. Toward this end, we define dynamic perceptual organization as an extension of the traditional (static) perceptual organization approach. Just as static perceptual organization groups coherent features in an image, dynamic perceptual organization groups coherent motions through an image sequence. Using dynamic perceptual organization, we propose a new paradigm for motion understanding and show why it can be done *independently* of the recovery of scene structure and scene motion. The paradigm starts with a spatiotemporal cube of image data and organizes the paths of points so that interactions between the paths and perceptual motions such as *common*, *relative* and *cyclic* are made explicit. The results of this can then be used for high-level motion recognition tasks.

1 Introduction

The analysis of dynamic image sequences is of fundamental importance for computational vision because of the generality of a moving observer and moving objects in the 3D world. Yet, image motion information has been used by researchers in computational vision primarily as one of several sources for reconstructing various intrinsic physical properties of the scene. Rather than focus on recovering quantitative "maps" of intrinsic 3D structure, an alternative approach is to emphasize the goal of perceptual organization, i.e., discovering primitive image relations that group sets of image features into relevant structures [1, 2, 3]. The key organizing principle is to find relations that are unlikely to have occurred by accident. For example, proximity, collinearity and parallelism have been used for organizing spatial data [3]. While Witkin and Tenenbaum suggested using the manner in which objects move as an organizing criterion [4], motion properties have not been used to their full potential in this respect.

As objects and contours move in the scene (i.e., undergo scene motion), their projections into the image also move, generating image or spatiotemporal motion. Over time, these projections sweep out spatiotemporal volumes, surfaces and curves. In this paper we propose that these spatiotemporal features are of fundamental importance for detecting perceptually-significant groupings of image features. We construct a hierarchical image-motion description as a means of recovering the spatiotemporal primitives that serve to distinguish physically-meaningful objects, processes and events. That is, the shape and interaction of spatiotemporal features provide sufficient information for object discrimination, interaction, and in many cases, identification and motion recognition.

Spatiotemporal features provide a powerful basis for organizing dynamic image data because they represent the complete spatial and temporal characteristics of the data. For example, depending on the complexity of the scene, the perceptual system can choose to use a sufficiently long interval of time and the spatiotemporal features therein so as to completely disambiguate between possible organizations. For example, in a scene containing widelyspaced, parallel dotted lines, an instantaneous temporal sequence (i.e., a single image) is sufficient to derive relevant feature groupings. A scene containing a few seemingly random points located on the spokes and rim of an invisible, rotating wheel, is interpretable only after a short period of time. The motion of a few points attached to a walking human will usually require a relatively long temporal sequence ("one or two steps" as observed by Johansson [5]). Because the necessary interval of time to organize the data is scene dependent, it is important to analyze long image sequences (i.e., on the order of hundreds of frames) in order to handle all situations.

Generalizing previous approaches to perceptual organization, the emphasis here is on discovering groupings of spatiotemporal image features that reflect meaningful structure of the scene. Specific 3D spatiotemporal (x-y-t) relations lead to inferences about 3D spatial (x-y-z) structure without requiring an explicit computation of scene surfaces or their motion in the scene. The reconstruction of quantitative 3D surface structures such as depth maps should be done, and only if necessary, *independently* of this stage. In situations where high-level object models are given, for example of a human body, our motion description and organization hierarchy will also enable recognizing coordinated sequences of events such as walking and throwing without ever recovering 3D scene structure or recognizing object parts. This is consistent with psychophysical evidence using moving light displays [5].

Historically, image motion has been used in two main ways: (1) to recover structure in the 3D scene, and (2) to recover instantaneous motion in the scene. Scene structure is often represented as the $2\frac{1}{2}$-D sketch or intrinsic images. Following the recovery of the physical structure of the scene, object recognition is performed by matching the derived 3D scene structure with 3D object models. Scene motion computes the motion of points in world coordinates. Only after the objects are recognized are the scene dynamics analyzed by using the scene motion to compute what, if any, high-level motion the objects are undergoing (e.g., walking by a human) [6, 7, 8]. This paradigm, which is the Marr approach with motion aspects emphasized, is shown in Figure 1. Note that the recognition of object motion is

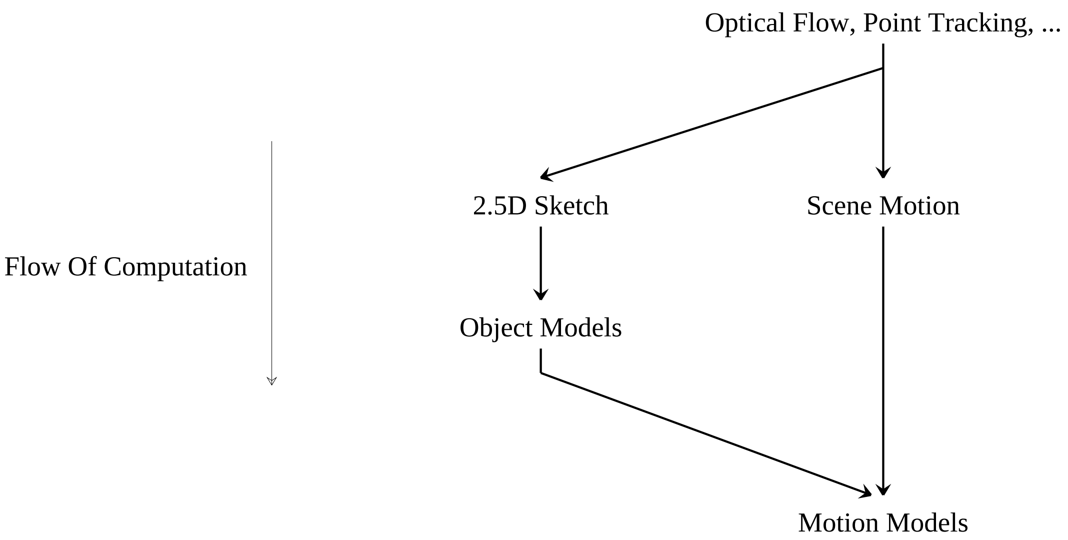


Figure 1: The traditional flow of computation. Note that motion information is not used up through object recognition, and *Motion Models* are computed last, after the objects have already been recognized. This figure shows only the motion aspects of a vision system, i.e., structure-from-stereo, for example, is not shown. Note that this figure is organized with structure on the left and motion on the right.

computed last.

Alternatively, Lowe and others have argued convincingly for the importance of bottom-up grouping of image features that can be used directly for recognition. This paper extends that approach to include dynamic perceptual organization as an important component of the process of organizing image features and inferring 3D structure. In some cases, the dynamic perceptual organization paradigm alone is sufficient for recognition as shown by the human visual system's ability to recognize certain objects and their motion [5]. Also, in cases where depth information is unavailable and perceptual organization of static image features is ambiguous, grouping dynamic features based on their spatiotemporal characteristics can be used for recognizing generic high-level motions such as rolling and swinging. As in the traditional perceptual organization paradigm, object recognition does not require prior recovery of scene structure. Hence we see our work to be a natural extension of previous work on perceptual organization. Figure 2 shows graphically our approach, given as an extension

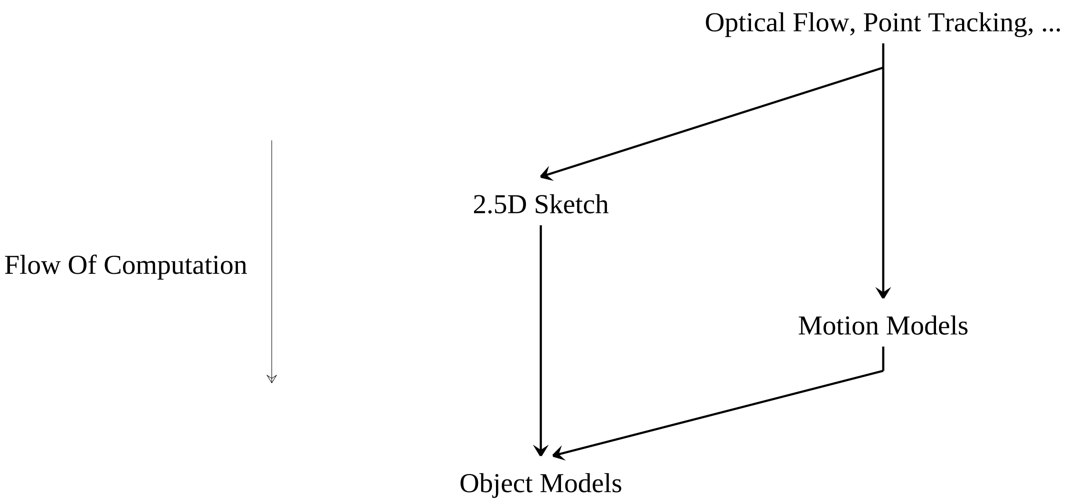


Figure 2: The dynamic perceptual organization paradigm. Note that *Motion Models* are recovered prior to the recovery of *Object Models* and independently of *Scene Structure*. In this paradigm the results of *Motion Models* can be used to aid in the recovery of *Object Models*. This figure shows only the motion aspects of a vision system, i.e., structure-from-stereo, for example, is not shown. Note that this figure is organized with structure on the left and motion on the right.

of the perceptual organization paradigm to emphasize image-sequence motion features.

The novel aspect of this work is that motion recognition is performed prior to object recognition and prior to recovery of scene structure, and does not require recovery of scene motion. Aloimonos [9] made a similar point, that scene structure is not necessarily needed in order to perform certain visual tasks. However, Aloimonos' justification was based on the benefits of a purposive vision system while ours is based on the fact that scene structure is not necessary for the recognition of high-level motions because, in general, the image-sequence motion uniquely identifies the high-level scene motion. That is, there is a one-to-one correspondence between image-sequence motion and scene motion. Hence, the problem of determining scene motion from image-sequence motion is well-posed (see Section 4).

In this paper we define a hierarchy of spatiotemporal structures (see Figure 3) that forms the dynamic perceptual organization paradigm. An image sequence is represented as a spatiotemporal cube of image data. The first level of the paradigm makes instantaneous motion of points in the cube explicit. The instantaneous motion is then organized into paths of points through an image sequence. These paths are then grouped such that each group represents a single coherent motion such as translation. Finally, the spatiotemporal paths of points can be organized so that perceptual organizations of motion such as *common* and *relative* motion are made explicit. Evidence for the existence of each of these levels in the human visual system can be found in the psychophysical literature and will be presented below. The primary purpose of this paper is to present this new paradigm and justify it from both the psychophysical and computational points of view.

The next section presents related work, showing how this approach has been largely ignored in previous methods. Next, the details of the dynamic perceptual organization paradigm are presented. As stated above, the novel aspect of this paradigm is that motion recognition can be performed much earlier in the computation hierarchy than has previously been done. Section 4 provides a computational argument why this is true. Section 5 presents our current results using this approach. Future directions for exploring the full power of this new paradigm are discussed in Section 6.

2 Related Work

Considerable psychophysical evidence exists supporting spatiotemporal features for perceptual organization and object recognition. Johansson's *moving light displays* (MLDs) are image sequences containing the motion of a few points of light, produced, for example, by attaching lights to the joints of a person walking in the dark [5]. Because each image consists of a few random points of light containing no structure, the perceptual organization of these points and the recognition of the objects is strictly due to the relations between the spatiotemporal curves swept out by the lights.

Johansson [5] used MLDs consisting of ten lights to examine human performance at interpreting articulated objects. MLDs were used so that only point motion information at the joints was available. In the first demonstration, adults watched an MLD of human walking, as viewed from the side of the walker. Johansson found that humans can very easily recognize the MLDs of walking. The second demonstration used moving light displays that were made with the walker moving toward the camera in directions between 80 and 45 degrees to the camera. The results with this display were similar to the results in the first demonstration. Johansson also tested displays for running, cycling, climbing, dancing in couples, various types of gymnastic motion, and others. In all cases adults correctly identified the motion.

Runeson [10] made MLDs that had points of light not only on the joints of a person but also at the corners of a box that the person was lifting. Observers of this MLD were able to recognize and judge the weight of the box. In this display the motion of the person and the box, and not the structure of the objects, were used to judge the weight. The structure could not have been used since the person's structure remained unchanged between displays; it was the motion of the person that varied with the weight of the box. Other studies [11, 12] found evidence that adults are capable of recognizing friends and the gender of a person from only the joint motions in MLDs.

Spelke also showed the importance of motion on object recognition, observing that young infants fail to recognize objects based on static configurational properties, but do apprehend objects by analyzing the motion of cohesive, bounded and spatiotemporally-continuous surfaces [13].

Related work on computational visual motion can be classified into four categories according to the type of information recovered from an image sequence. The four categories are: low-level structure, high-level structure, low-level motion and high-level motion. Low-level structure consists primarily of work that recovers 3D scene structure of rigid parts from image sequences (e.g., structure from motion). High-level structure has dealt with recovering and representing articulated objects [6, 7, 8]. However, given some object in an image sequence, all these articulated object representations are better suited to match the scene structure with a model rather than match the image motion or scene motion with a model.

Low-level motion has been primarily concerned with combining image sequences so that viewer-centered changes between frames due to motion in the scene is made explicit (e.g., optical flow). Work on high-level motion is concerned with organizing lower-level motion descriptions. The *trajectory primal sketch* of Gould and Shah [14] does this by organizing the image motion of points into translation, rotation and cycloidal primitive types. Goddard [15] organized a sequence of angular velocity changes of joints into a sequence that represented high-level motions such as walking. Yamato, Ohya and Ishii [16] were able to distinguish a small collection of motions under very controlled conditions by examining the variation of a measure based on the number of black pixels in a thresholded image sequence.

Very little previous work has addressed the issue of computing high-level motion descriptions prior to recovering scene structure or scene motion. Most of the work that might be considered to address this issue is actually recovering high-level structure and not modelling how objects move in order to recognize their motion [6, 7, 8]. Only the work of Gould and Shah [14] and Goddard [15] has addressed the issue of robustly interpreting the high-level motion through an image sequence.

3 The Dynamic Perceptual Organization Paradigm

The hypothesis of this paper is that it is possible to perceptually organize the image-sequence motion *independently* of the recovery of scene structure and scene motion. As pointed out in Section 2, all previous computational models first recover rigid parts and then use this information to index into the models of objects to be recognized.

A small number of researchers have argued that it is possible to recognize an object and its high-level motion without first recovering its high-level geometric structure [17, 16, 18, 14, 15]. The image-sequence motion of a few representative points rather than the low-level structure, e.g., rigidity, can be used to recognize an object [14]. Most of these

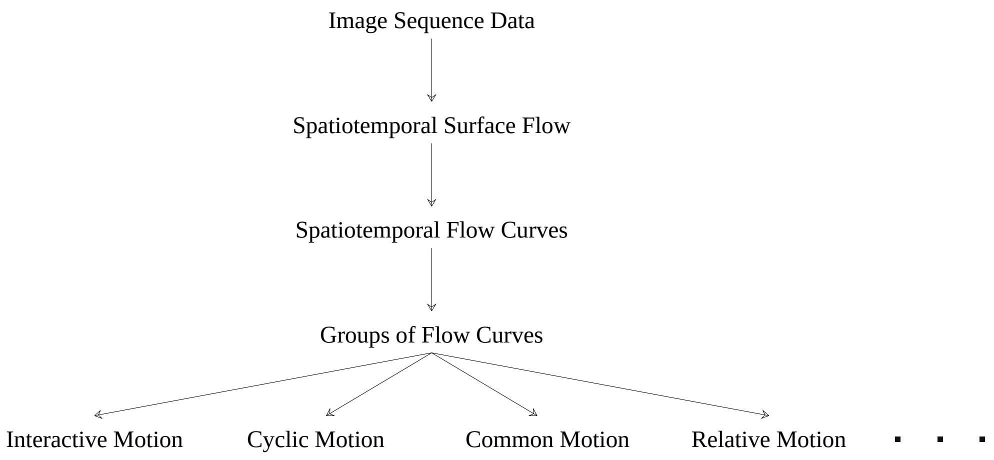


Figure 3: A hierarchy of dynamic perceptual groups. Spatiotemporal flow curves group spatiotemporal surface flow vectors into the paths of points through an image sequence. Similarly shaped flow curves are then grouped together. Specific perceptually-salient features are then detected from groups of flow curves. These perceptual features include, but are not limited to: interactive motion between groups (e.g, occlusion and disocclusion), cyclic motion, and common and relative motion.

approaches require high-level models that are rich in temporal information [15]. For example, the fundamental event in one of Goddard's *scenarios* is a change in angular velocity. A scenario becomes active when the correct sequence of these changes occur. In other words, the motions of joints is used to recognize, or index into, high-level motion models. 3D geometric structure had no role in this indexing.

Our goal is to organize the motion starting from the given image sequence data. We propose that motion should be perceptually organized in a hierarchical representation. We will construct this hierarchy non-purposively, i.e., without depending on high-level models, analogous to the way Marr developed the primal sketch and $2\frac{1}{2}$ sketch [19].

Figure 3 shows our hierarchy of dynamic perceptual groups derived from an image sequence. This hierarchy includes only the data-driven, i.e., non model-based, aspects. The image sequence data is given as a three-dimensional spatiotemporal (ST) volume, i.e., an image plane $\times$ time cube of pixels, and the fixed temporal depth of the cube provides a moving temporal window into the infinite "stream" of images. A first step in understanding the motion in an ST cube is to determine the instantaneous motion of each point in the cube, which we call the *spatiotemporal surface flow* [20]. This is the lowest level in our hierarchy. The next step is to organize the instantaneous motion of points into the long-range motion of each point through the ST cube up to the most recent frame. As additional images become available, the paths of points are extended into the new frames. An ST curve through an ST cube such that the tangent vector at a point on the curve equals the ST surface flow at that point is called a *spatiotemporal flow curve* [21]. This is the next level in our hierarchy.

Similarly shaped ST flow curves are then organized into groups so that each group represents a single coherent motion such as translation or rotation. From these groups even higher-level spatiotemporal structures can be organized. For example, spatiotemporal interaction between two groups of ST flow curves implies occlusion and disocclusion is occurring [21]. Periodic flow curves indicate cyclic motion (common among ambulatory objects) [22]. In addition, the ST flow curves in a group can be decomposed into the relative motion and common motion of the points. These two motions are equivalent to the relative and common motions that are observable by the human visual system [23].

As stated above, this hierarchy includes only bottom-up, i.e., non model-based, aspects. The results of the final level of this hierarchy, i.e., the different types of motion, can be used for motion and object determination from a database of models that explicitly incorporates the characteristic motions that are integrally associated with an object. These models can be of "simple" objects and motions such as a rolling wheel, or they can be more complex, such as an ambulatory biological object undergoing walking. Other examples of motion-based object models might include galloping horse, trotting horse, and billowing cloud.

The next two subsections give more details on ST flow curves, how they are recovered, organized, and interact. Sections 3.3 and 3.4 discuss the importance of relative and common motion, and why they are necessary. The importance of cyclic motion will not be discussed further in this paper since it has been described elsewhere [22]. Briefly, it is a significant generic spatiotemporal feature because it is characteristic of many natural objects and their motions. Further, its detection using ST flow curves demonstrates the representational power of ST flow curves.

3.1 Recovering Spatiotemporal Flow Curves

As described in Section 1, different types of motion require different temporal lengths before they can be perceptually grouped. Since an ST cube, constructed by stacking a long sequence of temporally-close images together, represents image-sequence motion over an arbitrarily long interval, it is the appropriate structure from which to begin our analysis.

A first step in understanding the motion in an ST cube is to determine the instantaneous motion of each point in the cube, called the *ST surface flow*. Assuming without loss of generality that the time between successive images is unity, ST surface flow is defined as a field of unit vectors $\mathbf{F} : \Re^3 \rightarrow \Re^3$, $\mathbf{F}(\mathbf{x}) = (\Delta x, \Delta y, \Delta t)$ which can be thought of as pointing in the direction that a point moves from one frame to the next. The temporal component of a vector indicates the speed, as the speed approaches 0, Δt approaches 1, and as the speed approaches infinity, Δt approaches 0.

ST surface flow can be computed as shown in [20] or traditional optical flow fields can be converted into ST surface flow fields. Vectors in an optical flow fields are typically two-dimensional, $\mathbf{V} = (\Delta x,\ \Delta y)$, with Δx and Δy indicating the direction of motion and the length, $\|\mathbf{V}\|$, indicating the speed. An optical flow vector, $\mathbf{V}$, is converted to an ST surface flow vector, $\mathbf{V^{st}}$, by adding a temporal component which varies from 1 to 0 as $\|\mathbf{V}\|$ varies from 0 to infinity. $1 - \frac{\tan^{-1}(\|\mathbf{V}\|)}{\pi/2}$ is just such a quantity. The resulting vector is then made a unit vector. So

$$\mathbf{V^{st}} = \frac{\left(\Delta x,\ \Delta y,\ 1 - \frac{\tan^{-1}(\|\mathbf{V}\|)}{\pi/2}\right)}{\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2}}$$

Once the ST surface flow is computed, the next step is to organize the instantaneous motion of points into the long-range motion of each point through the ST cube up to the most recent frame. Given the ST surface flow over many frames, *ST flow curves* through the ST cube are then recovered. Loosely, flow curves are started in the first frame and flow through the cube such that the tangent vector at a point on a flow curve equals the ST surface flow at that point. As additional images become available, these curves are extended into the new frames.

An ST flow curve can be represented as a parameterized space curve. A parameterized ST 3D curve, α(t), is a map $\alpha : \mathbf{I} \rightarrow \Re^3$ of an open interval $\mathbf{I}$ = (a, b) of the real line $\Re$ into $\Re^3$. α defines a correspondence which maps each $t \in \mathbf{I}$ into a point $\alpha(t) = (x(t), y(t), t) \in \Re^3$.

Using prime to denote the partial derivative with respect to t, the vector

$$(x'(t), y'(t), 1) = \alpha'(t) \in \Re^3$$

is called the tangent or velocity vector of the curve α at t.

Assume for the moment that the ST surface flow is smooth. Given a smooth ST surface flow, an ST flow curve α is defined such that the velocity vector of α at each point equals the ST surface flow at that point. (See Figure 4.) Recall that the ST surface flow is defined by

$$\mathbf{F} : \Re^3 \rightarrow \Re^3 \qquad \mathbf{F}(\mathbf{x}) = (\Delta x, \Delta y, \Delta t)$$

Requiring the velocity of α to equal $\mathbf{F}$ is equivalent to

$$\alpha'(t) = \mathbf{F}(\alpha(t)) \tag{1}$$

A given flow curve has the initial condition $\alpha_i(0) = (x_0, y_0, 0)$, where (x_0, y_0) are the coordinates of the pixel in the first frame. Using coordinates (x, y, t) for the ST cube, Eq. (1) can be rewritten as the simultaneous equations

$$\begin{aligned} x'(t) &= F_1(x(t), y(t), t) \\ y'(t) &= F_2(x(t), y(t), t) \\ 1 &= F_3(x(t), y(t), t) \end{aligned}$$

with initial conditions

$$(x(0), y(0), t) = (x_0, y_0, 0)$$

where $\mathbf{F} = (F_1, F_2, F_3)$.

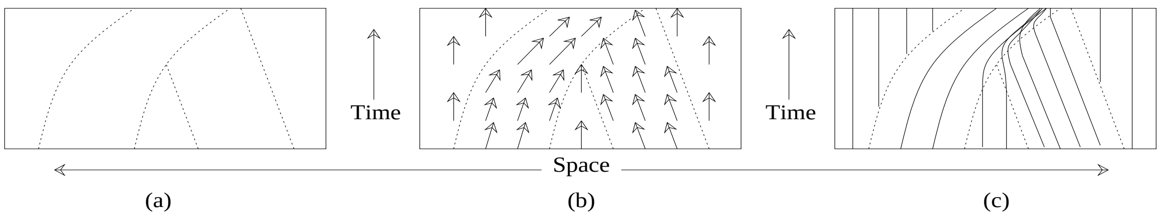


Figure 4: (a) A spatiotemporal image with two objects, one accelerating right and one translating left. (b) The ST surface flow. (c) ST flow curves.

Many methods can be used to solve this system of equations given the starting points in the first frame and the ST surface flow, **F**, defined at every pixel in the ST cube. However, since **F** is only defined at coordinate points and must be interpolated at intermediate pixels, the relatively simple Runge-Kutta method [24, 25] is appropriate. More sophisticated methods where the increment in t varies depending upon the complexity of **F** are not used since there is no reason not to use the smallest increment in t available, namely 1.

3.2 Grouping Spatiotemporal Flow Curves

Once the flow curves are recovered, similarly shaped ST flow curves are organized or grouped so that each group represents a single coherent motion such as translation or rotation. Figure 4 shows an example of how the shapes of ST flow curves can be used for the interpretation of a long image sequence. Figure 4(a) shows a slice from an ST cube with one spatial and one temporal dimension. There are two objects, one accelerating to the right and one translating to the left. The ST surface flow is shown in Figure 4(b). As expected, the flow for the left object points toward the right, the flow for the right object points toward the left, and the flow for the background points straight into time. Figure 4(c) shows the resulting ST flow curves.

The flow curves for the left object all have similar shape, as do the flow curves for the right object. These curves can be grouped using properties of curves that measure the shape of the curve, e.g., curvature and torsion. However, the flow curves for the right object and the background are initially straight so have identical curvature and torsion. So, in addition to examining the shape, the slope of the curve is also required. By applying standard clustering techniques using curvature, torsion and slope, three distinct groups are formed. Each group represents the motion of a single object or the background over a arbitrarily long period of time.

Once the flow curves are recovered, it is relatively straightforward to compute the shape describing properties such as curvature and slope. The results of the Runge-Kutta method give a sequence of points that define each flow curve. In order to compute shape-description properties of flow curves a quadratic curve is fit to each set of points. A separate curve segment, centered at each point, is fit for every point making up a flow curve. This is done using a 1D version of the quadratic surface fitting procedure described by Besl and Jain [26]. Once the quadratic curve segment is fit, the partial derivatives at a point can be recovered and the curvature computed using the following equation:

$$\kappa = \frac{\sqrt{A^2 + B^2 + C^2}}{\left((x')^2 + (y')^2 + (t')^2\right)^{3/2}}$$

where

$$A = \begin{vmatrix} y' & t' \\ y'' & t'' \end{vmatrix} \qquad B = \begin{vmatrix} t' & x' \\ t'' & x'' \end{vmatrix} \qquad C = \begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}$$

By using ST flow curves as the primitive features of motion description, the image-sequence motion over an arbitrarily long time is explicitly represented in a concise, well-defined form. Further, similarly shaped ST flow curves, i.e., curves with similar curvature and slope, can be organized into groups using standard clustering techniques and interactions between flow curves (Interactive Motion in Figure 3) can be detected and interpreted. (It was argued by Allmen and Dyer [21] that torsion is not necessary to distinguish between different types of motion.) Specifically, a hierarchical clustering algorithm is used to initially cluster the flow curves. Each resulting group then represents a single coherent motion. As curves change shape, because the object they are associated with becomes occluded for example, clusters of curves will merge and split. Specifically, K-Means clustering is used to update the clusters at each subsequent time step. K-Means works in an iterative manner, updating the clustering as each flow curve is extended into the current frame. That is, for each flow curve, using a fixed width interval ending at the current time, the similarity of curvature and slope values of the flow curve and its cluster's mean is computed. If this similarity is less than the similarity of the flow curve and another cluster's mean, the flow curve is moved into the other cluster. From these groups of ST flow curves, separate moving objects can be hypothesized, and occlusion and disocclusion between them can be identified by examining how groups merge and split [21]. Results of this process are shown in Section 5.

Using the groups of ST flow curves, we can now organize the motion into higher-level representations such as the common and relative motion within and between these groups of ST flow curves. The next section defines these different types of motion and gives examples to show their importance in higher levels of motion representation.

3.3 Common and Relative Motion

The human visual system (HVS) perceives a rolling wheel as translating across the field of view and rotating about the center of the wheel. Even though the absolute image-sequence motion of a point on the rim of the wheel is a cycloid, the HVS does not perceive this cycloid. Instead, the HVS perceives a decomposition of the motion into two components: the *common motion* of translation and the *relative motion* of rotation about the center of the wheel. The cycloidal *absolute motion* is generally not perceived.

Even though the absolute cycloidal motion is not perceived, the following equation holds [23]:

$$\text{common motion} \quad + \quad \text{relative motion} \quad = \quad \text{absolute motion} \tag{2}$$

For example, we can show how this equation holds for a rolling wheel. Recall that a parameterized curve in $\Re^2$ is a map α: $\mathrm{I} \rightarrow \Re^2$ of an open interval $\mathrm{I} = (\mathrm{a}, \mathrm{b})$ of the real line $\Re$ into $\Re^2$. Let $C(t)$, $R(t)$ and $A(t)$ parameterize the path of the common, relative and absolute

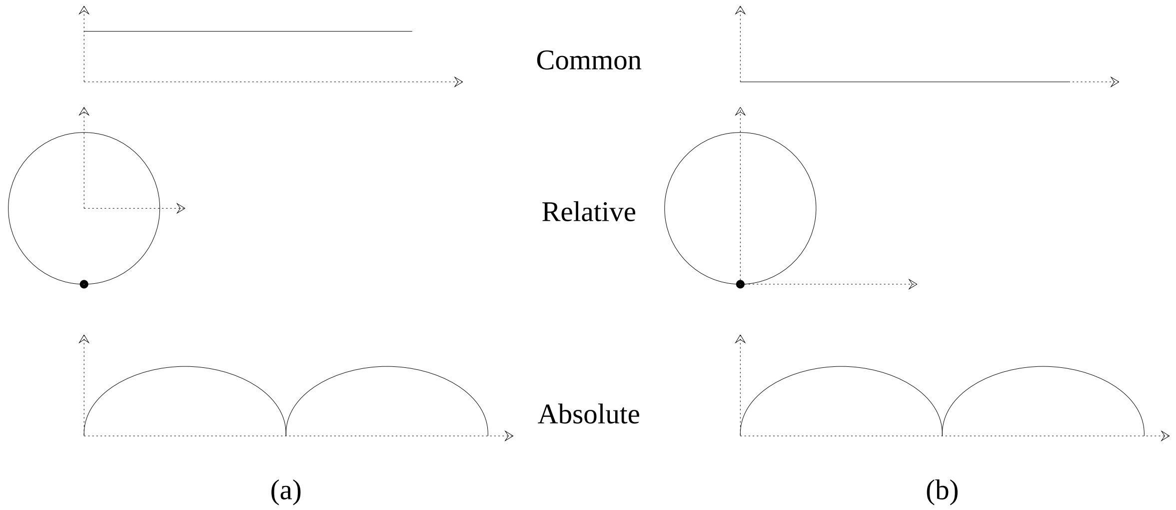


Figure 5: Two possible perceptions of common, relative and absolute motions for a point on the rim of a rolling wheel.

motions, respectively. For a rolling wheel we have:

$$\begin{aligned} C(t) &= (t, 1) \\ R(t) &= (-\sin(t), -\cos(t)) \end{aligned}$$

Adding $C(t)$ and $R(t)$ we have:

$$\begin{aligned} A(t) &= C(t) + R(t) \\ &= (t - \sin(t), 1 - \cos(t)) \end{aligned}$$

which parameterizes a cycloid. Figure 5(a) shows the three curves. Note that this is only one solution to Eq. (2). Another solution is obtained by vertically translating $C(t)$ and $R(t)$, as shown in Figure 5(b). Another, more obscure, solution is shown in Figure 6. This solution is generally not perceived by the HVS even though the relative and common motions still combine to equal the absolute motion.

There is no unique solution for the relative and common motions because there is only one equation, Eq. (2), and two unknows, R and C. The goal is to find a solution that has the "simplest" relative and common motions, i.e., a solution that is similar to the solution shown in Figure 5 rather than the solution shown in Figure 6.

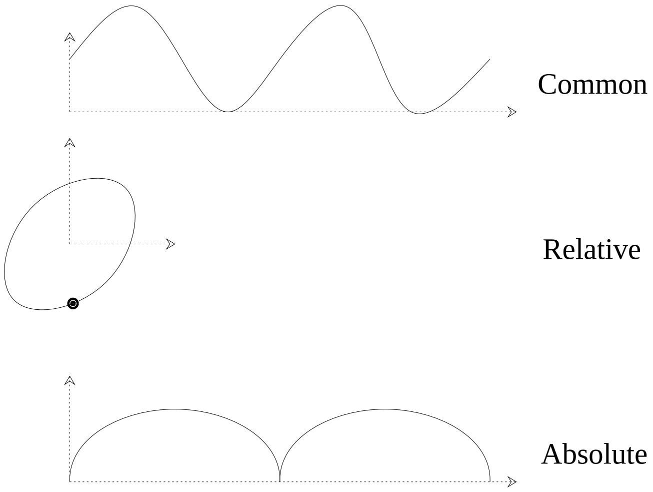


Figure 6: An unlikely perception of common, relative and absolute motions for a point on the rim of a rolling wheel.

In general, the absolute motion is the motion once an arbitrary viewer-relative frame of reference is specified. The common motion is the motion of the whole configuration relative to the viewer. The relative motion is the motion relative to another point. These loose definitions of relative and common motions are based on how the HVS perceives motion. There are no “correct” solutions to Eq. (2), only solutions that are similar to the HVS’s perception. Therefore, how the HVS perceives these motions can be used to derive additional constraints to solve Eq. (2).

Psychologists have long recognized this perceptual phenomenon [27, 28, 29], but have failed to develop a complete model of how the HVS computes relative and common motion [5, 23]. One commonality in those models is to minimize the total complexity of the relative and/or common motions [30, 31, 23]. Minimizing the motions can be viewed as making it as simple as possible. Unfortunately, how and what is being minimized has been poorly defined.

The importance of common and relative motions in higher-level representations of motion becomes apparent when examining situations where the absolute motion is distorted. In the discussion below the relative motion is decomposed into a spatial component and a temporal component and the importance of relative motion becomes obvious.

3.4 Spatial and Temporal Relative Motion

The relative spatial locations of rigid parts is clearly important in a motion description because if the rigid parts are not positioned correctly relative to each other, the high-level motion is drastically altered. For example, consider an overhand throwing motion verses an underhand throwing motion. In both cases, the relative motions of the parts of the arm are similar. However, in the overhand case, the upper arm is spatially below the forearm whereas in the underhand case the upper arm is spatially above the forearm. In order to prevent the two types of throwing from being equivalent, relative spatial information must be incorporated.

Cutting showed that relative spatial information is a necessary component of MLDs in order for the HVS to perceive walking [32]. He presented subjects with a spatially anomalous "walking" MLD, where the initial positions of the lights were altered while keeping their individual absolute motions correct. None of the subjects detected the walking motion.

The relative times that rigid parts move is also of fundamental importance in a motion description since if the rigid parts do not move at the correct relative times, the interpretation of high-level motion is altered. For example, consider a walking horse verses a trotting horse. While walking, the foot sequence is: right hind, right front, left hind, left front. While trotting, the foot sequence is: left front and right hind together, right front and left hind together. The paths of the legs are similar in the two situations, but the relative times of the movements differ. In order to prevent the two types of running from being equivalent, relative temporal information must be incorporated.

Like the spatially anomalous "walking" MLD experiment described above, an experiment involving a temporally anomalous walking MLD could be performed. In this case, each light would undergo the correct motion, but at the incorrect time. For example, each light could start its motion at different times. It is expected that subjects would not detect the walking motion in this situation.

4 Uniqueness of Dynamic Perceptual Organization

Since an infinite number of scene configurations can give rise to any given image, the process of computing scene structure from a single image is ill-posed. A large amount of computer vision research has attempted to discover ways to overcome this problem. For example, using structure-from-motion or photometric stereo, one can compute the scene structure. The added constraints of how objects move through an image sequence or the effects of lighting changes, make it possible (with additional assumptions) to recover scene structure.

Alternatively, researchers have used perceptually-salient feature groupings in order to better interpret a single image. These groupings are chosen so that the probability is low that such a relation of features would accidently appear in an image. As described in Section 3, we use groupings of ST flow curves to organize the image-sequence motion. However, unlike static perceptual organization, dynamic perceptual organization is guaranteed that dynamic features in an image sequence do not arise by chance. That is, whereas static perceptual organization must estimate the likelihood that a group of features did not occur by accident, dynamic perceptual organization does not need to make this assumption. This results because of the one-to-one correspondence between the motion in an image sequence and the scene motion that generated it. Because of this uniqueness, we will see that the actual recovery of scene structure and scene motion is unnecessary and, furthermore, undesirable.

The most novel aspect of the dynamic perceptual organization paradigm is that scene structure and scene motion are not recovered. In this section we briefly review the work of Ullman [33], Hoffman and Flinchbaugh [34], and Bennett and Hoffman [35] to show the assumptions necessary to recover structure from motion. Given these assumptions about the scene, the image-sequence motion uniquely represents the scene motion. That is, unlike the case with a single image where an infinite number of scene configurations can give rise to, say, any given contour, a unique scene configuration over time can generate a given set of ST flow curves. This makes the task of inverting the ST flow curves into scene motion unnecessary since the flow curves already uniquely represent the scene motion. Because of this one-to-one mapping between scene motion and image-sequence motion, the image sequence can be used directly to organize the motion. Ultimately, the image-sequence motion can also be used for motion recognition. This differs from other static intrinsic features, such as the occluding contour, which do not uniquely determine the scene structure.

It is important to keep in mind that the structure-from-motion work presented below is only presented to show the uniqueness between scene structure and scene motion, and image-sequence motion. The fact that all the work presented requires only a few frames or viewpoints does not mean that dynamic perceptual organization requires only a few frames. As discussed in Section 1, different types of motion require different temporal lengths before they can be perceptually grouped. Perceptually grouping motions is different from computing the absolute motions of points in the scene. Nonetheless, as will be shown below, there is a one-to-one correspondence between scene motion and image-sequence motion.

The *rigidity assumption* states that any set of elements undergoing a spatiotemporal transformation and that has a unique interpretation as a rigid body moving in the scene, should be interpreted as such a body in motion. Ullman [33] used this assumption to show:

> Given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views is uniquely determined up to a reflection about the image plane.

The structure-from-motion computation varies depending upon whether orthographic or perspective projection is assumed, but in either case the scene structure is uniquely computable up to a reflection about the image plane. There are no restrictions on the spatiotemporal extent of this approach, so if more points or views are available, the scene structure can be computed more accurately.

The views used in Ullman's approach can be considered as existing in space-time. That is, each view is from some spatial position at some time. Since these points move over time, their projections into the image move over time, sweeping out ST flow curves. Therefore, given four ST flow curves of at least three frames in length, the scene location and scene motion of the four points is computable.

With the rigidity assumption and Ullman's result, there exists a one-to-one function from ST flow curves to scene structure. That is, the flow curves uniquely determine the scene structure. However, Ullman's result cannot explain the performance of the HVS when viewing MLDs since no rigid part has four lights on it. But by making additional assumptions, one can make similar statements about the motion of complex objects and provide a computational model for the HVS's performance at recognizing MLDs.

Since the rigidity assumption alone is not sufficient to explain the HVS's performance, Hoffman and Flinchbaugh [34] used the *planarity assumption* in order to reduce the number of points needed to recover scene structure. The planarity assumption states that any set of elements undergoing a 2D transformation that has a unique interpretation as a pairwise-rigid structure moving in one plane, should be interpreted as such a body in motion. Using this assumption they showed:

> Given two distinct orthographic projections of the three endpoints of two rigid rods linked in a hinge joint to form a pairwise-rigid structure which is constrained to move in one plane, the structure and motion compatible with the two views are uniquely determined (up to a reflection about the image plane).

Thus it is possible to explain the HVS's performance with MLDs if we believe that the HVS exploits the planarity assumption.

Bennett and Hoffman [35] made even stronger statements about recovering scene structure. By assuming that the axis of rotation stays fixed, called the *fixed axis assumption*, they showed:

> Given four orthographic projections of two points moving at independent angular velocities, the axis of rotation and the relative positions of the points in the scene

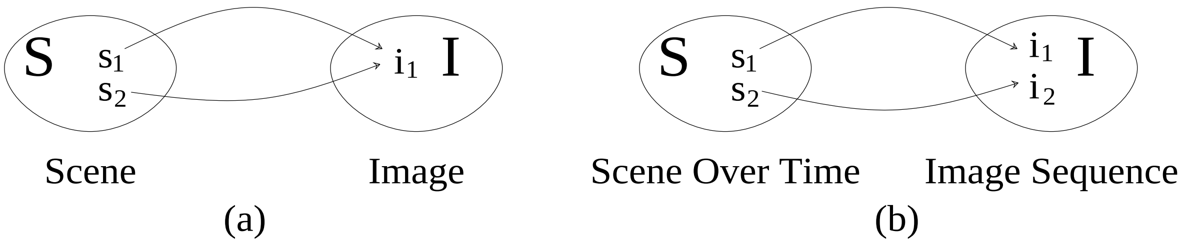


Figure 7: (a) The mapping from a scene to an image. $p : S \rightarrow I$. It must be assumed that the probability that $p(s_1) = p(s_2)$ is low. (b) The mapping from a scene over time to an image sequence. $p : S \rightarrow I$. It is known that probability that $p(s_1) = p(s_2)$ is zero.

are uniquely determined up to a reflection about the image plane.

In this case, not even the rigidity assumption is needed.

We will ignore for the moment that these structure-from-motion results give solutions only up to a reflection about the image plane. That is, we will ignore the fact that the computation of structure from motion is not unique and gives two possible scene structures. In the discussion below we will assume the computed structure and motion is unique.

The significance of these results is that there exists a one-to-one function from ST flow curves to scene structure and scene motion. That is, ST flow curves uniquely determine the scene structure and scene motion. Figure 7(a) shows the mapping, p, from the set of all scene configurations, S, to the set of all images, I. p represents the projection process from the scene to an image. Given a point $s \in S$, $p(s) \in I$ is the resulting image for a given viewpoint. It is well known that p is not 1-1. Therefore, the inverse of p does not exist. If one makes inferences about the scene based on a single image, it must assumed that the probability is low that more than one scene configuration could give rise to the same image. That is, given s_1 and $s_2 \in S$, it must be assumed that the probability that $p(s_1) = p(s_2)$ is low. Furthermore, it must be assumed that any conclusions made from $p(s_i) \in I$ about the scene refer to the more probable scene. That is, if s_1 is the most-likely scene configuration, it must be assumed that any conclusion made from $i_1 \in I$ actually refers to s_1 and not s_2.

Figure 7(b) shows the mapping from the set of all scene configuration over time, S, to the set of all image sequences, I. p represents the projection process from the scene to an image sequence. The results of this section show that p *is* invertible. So, for any $i \in I$, $p^{-1}(i)$ is uniquely defined. Therefore, it must no longer be hypothesized that any conclusion made from an image sequence refers to the true scene configuration. This is guaranteed to be the case because $p^{-1}(i)$ specifies only one scene configuration.

Perhaps even more important than being unnecessary, the recovery of scene structure and scene motion is also undesirable. This is because the process, even when well-posed, is unstable because of non-linear constraints [36]. In practice, since ST flow curves are noisy, using them or any recovered image motion representation will result in very noisy computed scene structure and scene motion. Therefore, rather than use this unstable process, we believe it should be avoided altogether. It is possible to avoid it because the image-sequence motion uniquely represents the scene.

Alternatively, one could use an "active observer" approach and make the problem stable [36], but, again, this is unnecessary for the reasons stated above.

5 Results

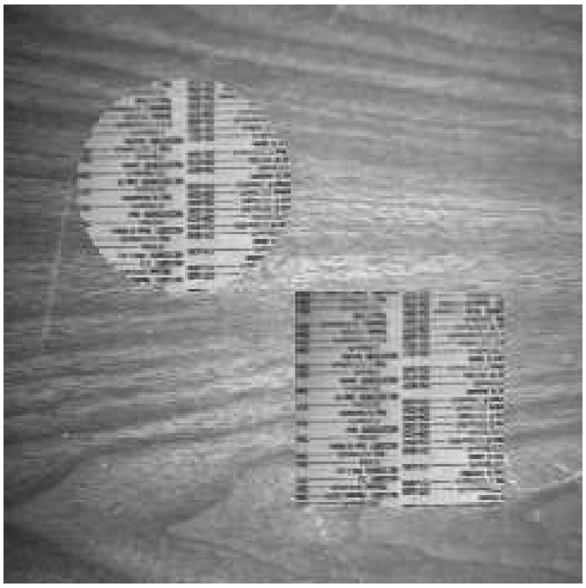
In this section we show computational results for the first steps of our dynamic perceptual organization approach. In particular, the computation of the instantaneous motion of points in an ST cube is presented, followed by the recovery and grouping of ST flow curves. Results such as these can then be used for further computations such as the computation of relative and common motion. See [20] for more examples of the computation of ST surface flow and see [21] for more results of computing ST flow curves.

Figure 8(a) shows one frame of a seven frame sequence of a phone book page rotating about a vertical line through the middle of the page at 0.01 radians/frame and a phone book page rotating about its center at 0.01 radians/frames. The image sequence was synthetically generated by transforming a planar region of a phone book page over a larger image of a table top. The image sequence was smoothed using a 3D Gaussian-weighted kernel with a standard deviation of 2 pixels.

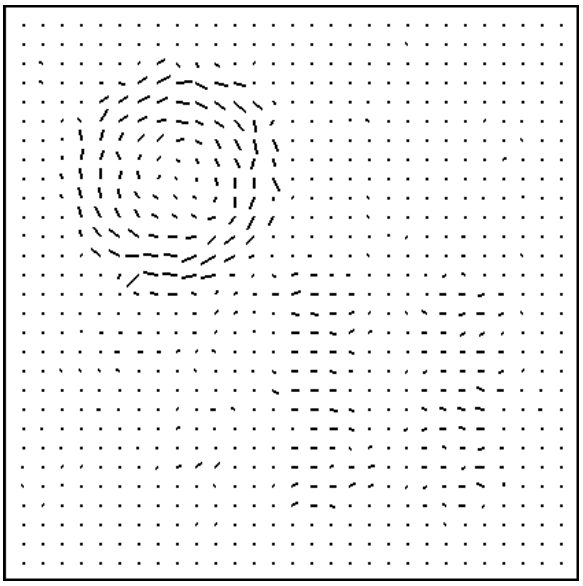
ST surface flow was computed at every fourth pixel in every frame. Figure 8(b) shows the resulting ST surface flow. A single temporal slice from the middle of the ST surface flow field is shown. Time is into the page so the vectors shown are the result of projecting the ST surface flow vectors into the spatial plane. A vector showing no motion, i.e., no spatial component, is oriented straight into time and appears as a dot. The greater the speed of a point, the greater the spatial component of the vector and hence, the longer the vector appears in the spatial projection. The resulting flow for each sequence was convolved with a $3 \times 3 \times 3$ median filter and smoothed using a 3D Gaussian-weighted kernel with standard deviation 0.66.

Figure 9 shows the resulting ST flow curves for the image sequence in Figure 8. The front and top views of each of the three detected groups of flow curves are shown. As expected, there is a group of flow curves for each object and for the background. The flow curves near the centers of rotation are almost straight and therefore were grouped with the background.

Figures 10 and 11 show results of computing Interactive Motion information (see Figure 3). Figure 10 shows the ST flow curves and their grouping from a 115 frame sequence of one square surface translating to the left in front of another square surface translating to the right. The top and front views of the four resulting groups are shown. Figure 11 shows the group after the sequence is extended to 170 frames. Initially, the flow curves associated with the surface translating to the right were in the center group since they were slanted toward the right, indicating translation to the right. But as this surface became occluded, the flow curves changed shape and became slanted to the left. At this point the flow curves were grouped with the group of flow curves associated with the occluding surface. This state of the motion grouping is shown in the figure.



(a)



(b)

Figure 8: (a) Middle frame from a 7 frame sequence of a phone book page rotating about a vertical line through the middle of the page at 0.01 radians/frame and a phone book page rotating about its center at 0.01 radians/frames. (b) ST surface flow for the middle frame. The final flow was smoothed using a 3×3×3 median filter.

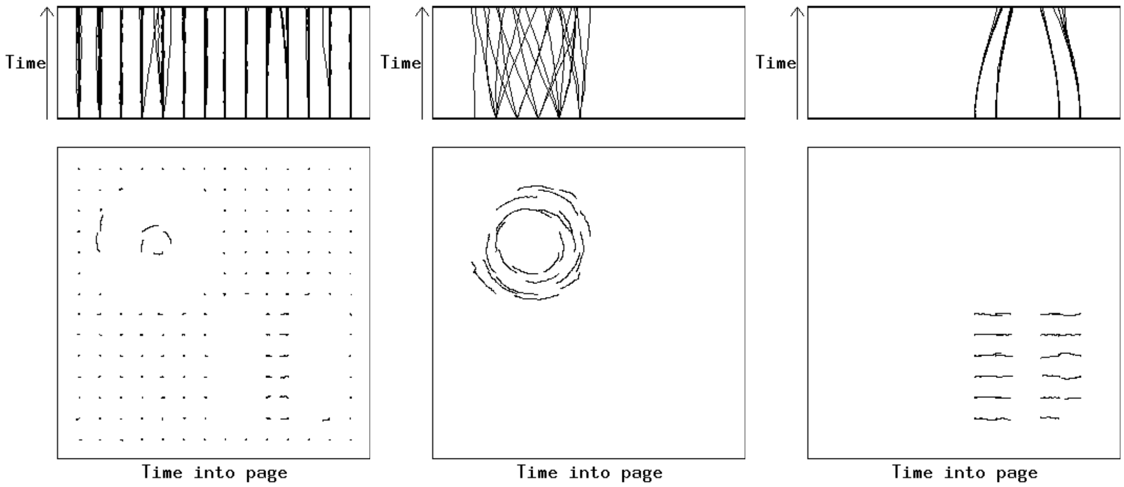


Figure 9: ST flow curves from a 115 frame sequence of a phone book page rotating about a vertical line through the middle of the page at 0.01 radians/frame and a phone book page rotating about its center at 0.01 radians/frames. The front and top views of the resulting groups are shown. The flow curves near the centers of rotation are almost straight and therefore were grouped with the background.

6 Concluding Remarks

In this paper we defined dynamic perceptual organization as an extension of the traditional (static) perceptual organization method. Just as static perceptual organization groups coherent features in an image, dynamic perceptual organization groups coherent motions through an image sequence. Using dynamic perceptual organization, we proposed a new paradigm for motion understanding and showed why it can be done *independently* of the recovery of scene structure and scene motion.

To date, the overwhelming use of motion in computational vision has been to recover the 3D structure in the scene. Our paradigm makes use of the fact that this is not the only way to approach image-sequence understanding. Rather than compute scene structure, we show why it is unnecessary to do so. Our approach does not rule out this more traditional approach, but it does have advantages over the traditional approach—the unstable computation of structure-from-motion can be avoided.

Our approach to image-sequence understanding can be used for various motion problems

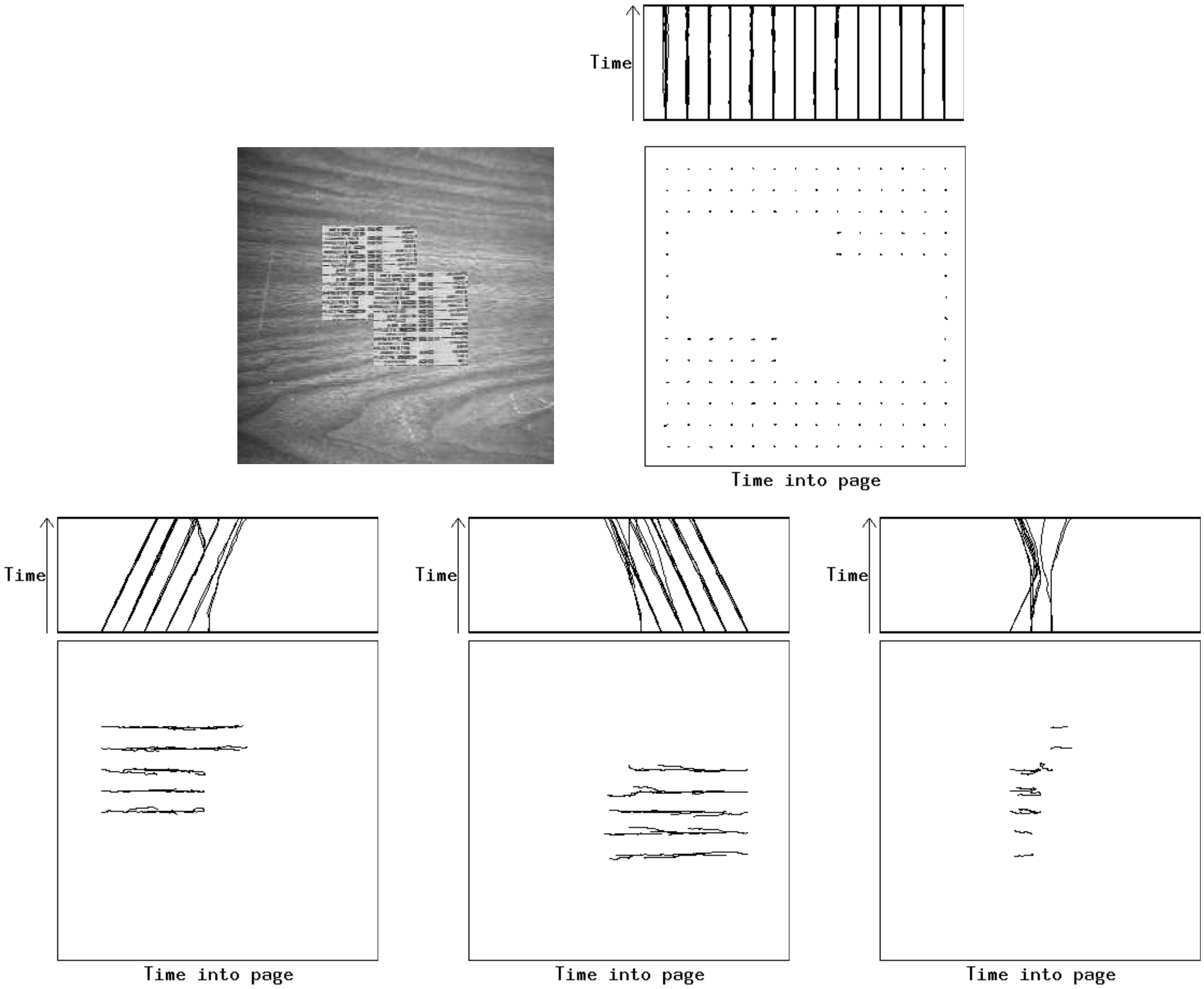


Figure 10: ST flow curves from a 115 frame sequence of two phone book pages, one translating left at 0.4 pixels/frame and partially occluding the other, which is translating right at 0.4 pixels/frame. One frame of the sequence is shown on the left. The front and top views of the resulting four groups are shown. A fourth group results in the area where the two pages overlap because the flow curves in this area are shaped like flow curves generated by the left page for a while then shaped like the flow curves generated by the right object. This results in the flow curves shaped like no other flow curves so a fourth group remains. As the flow curves were extended into subsequent frames, the flow curves in the extra group became shaped more like the flow curves following the occluding object. K-Means then moved the flow curves into the group associated with the occluding object. This is shown in Figure 11.

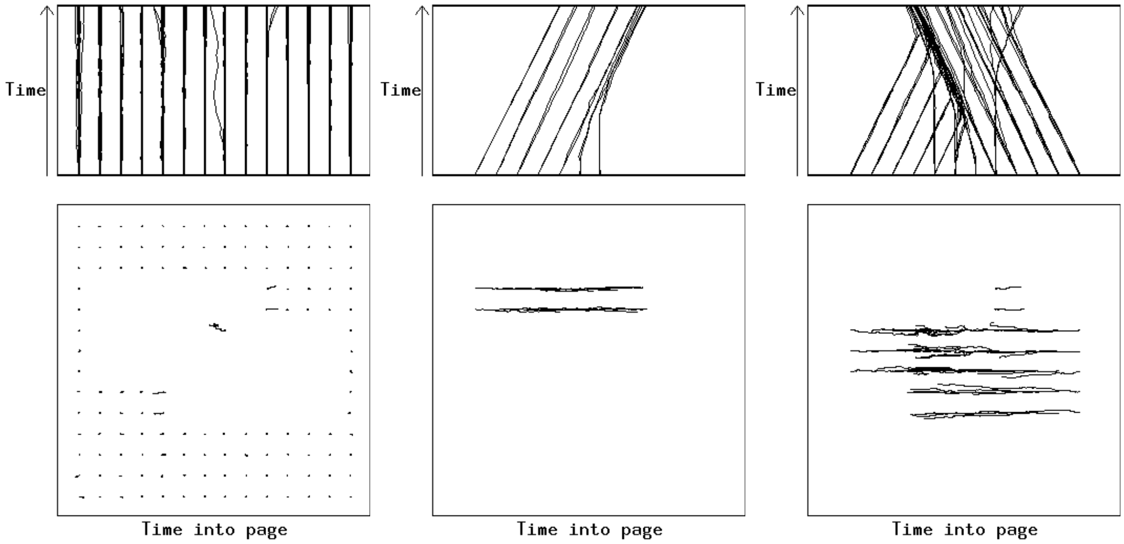


Figure 11: ST flow curves from Figure 10 extended to 170 frames. The top and front views of the three groups resulting from K-Means are shown. The number of flow curves in the fourth group was low enough that it was deleted. All of the flow curves that were generated by the lower part of the object translating right have merged into the group associated with the object translating left. Also, flow curves generated from the background have merged into the groups for the translating objects. There are two flow curves in the right group that should be in the center group. These errors are corrected by K-Means after the flow curves are extended a few more frames.

and can be incorporated into complete vision systems for analyzing both model-based and non-model-based motions of non-rigid objects. Minimally, our paradigm provides a rich organization of the motion in an image sequence. In a model-based system our organization can be used for motion understanding and motion recognition.

We showed promising results for the early levels of our approach, indicating that the computation of spatiotemporal features and their relations is an important and well-defined step. These results and other work on cyclic motion detection [22] show that it is possible to organize the motion through an image sequence independently of scene structure and scene motion. Future work will continue to explore the full potential of this dynamic perceptual organization paradigm.

References

[1] A. P. Witkin and J. M. Tenenbaum, "On the role of structure in vision," in *Human and Machine Vision* (J. Beck, B. Hope, and A. Rosenfeld, eds.), 481–543, Academic Press, New York, 1983.

[2] D. G. Lowe and T. O. Binford, "Perceptual organization as a basis for visual recognition," in *Proc. American Assoc. Artificial Intelligence*, pp. 255–260, 1983.

[3] D. G. Lowe, *Perceptual Organization and Visual Recognition.* Kluwer, Boston, 1985.

[4] A. P. Witkin, "Scale-space filtering," in *Proceedings Int. Joint Conf. Artificial Intelligence*, pp. 1019–1021, 1983.

[5] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, 14, 1973, 201–211.

[6] D. Marr and L. Vaina, "Representation and recognition of the movements of shapes," *MIT AI Memo 597*, October, 1980.

[7] D. Hogg, "Model-based vision: A program to see a walking person," *Image and Vision Computing*, 1, 1983, 5–20.

[8] K. Akita, "Image sequence analysis of real world human motion," *Pattern Recognition*, 17, 1984, 73–83.

[9] J. Aloimonos, "Purposive and qualitative active vision," in *Int. Conf. Pattern Recognition*, pp. 346–360, 1990.

[10] S. Runeson and G. Frykholm, "Visual perception of lifted weight," *Journal of Experimental Psychology: Human Perception and Performance*, 7, No. 4, 1981, 733–740.

[11] J. Cutting and D. Proffitt, "Gait perception as an example of how we may perceive events," in *Intersensory perception and sensory integration* (R. Walk and H. L. Pick, eds.), 249–273, New York: Plenum, 1981.

[12] L. MacArther and R. Baron, "Toward an ecological theory of social perception," *Psychological Review*, 90, 1983, 215–238.

[13] E. S. Spelke, "Origins of visual knowledge," in *Visual Cognition and Action, Vol. 2* (D. Osherson, S. Kosslyn, and J. Hollerbach, eds.), 99–127, MIT Press, Cambridge, Mass., 1990.

[14] K. Gould and M. Shah, "The trajectory primal sketch: A multi-scale scheme for representing motion characteristics," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 79–85, 1989.

[15] N. H. Goddard, "Recognizing animal motion," in *Proc. of Image Understanding Workshop*, pp. 938–944, 1988.

[16] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Computer Vision and Pattern Recognition Conf.*, pp. 379–385, 1992.

[17] R. Polana and R. C. Nelson, "Recognition of motion from temporal texture," in *Computer Vision and Pattern Recognition Conf.*, pp. 129–134, 1992.

[18] M. Allmen, *Image sequence description using spatiotemporal flow curves: Toward motion-based recognition.* PhD thesis, University of Wisconsin-Madison, 1991. (Available as Computer Sciences Department Technical Report #1040).

[19] D. Marr, *Vision.* San Francisco: Freeman, 1982.

[20] M. Allmen and C. R. Dyer, "Computing spatiotemporal surface flow," in *Proc. 3rd Int. Conf. Computer Vision*, pp. 47–50, 1990.

[21] M. Allmen and C. R. Dyer, "Long-range spatiotemporal motion understanding using spatiotemporal flow curves," in *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 303–309, 1991.

[22] M. Allmen and C. R. Dyer, "Cyclic motion detection using spatiotemporal surfaces and curves," in *Proc. 10th Int. Conf. Pattern. Recognition*, pp. 365–370, 1990.

[23] J. E. Cutting and D. R. Proffitt, "The minimum principle and the perception of absolute, common, and relative motions," *Cognitive Psychology*, 14, 1982, 211–246.

[24] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C.* Cambridge University Press, 1988.

[25] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, 1971.

[26] P. J. Besl and R. C. Jain, "Invariant surface characteristics for 3D object recognition in range images," *Comp. Vision, Graphics, and Image Proc.*, 33, 1986, 33–80.

[27] G. Johansson, *Configurations in event perception.* Almqvist & Wiksell, 1950.

[28] J. Hochberg, "Effects of the gestalt revolution: The Cornell symposium on perception," *Psychological Review*, 64, 1957, 73–84.

[29] H. Wallach, "Visual perception of motion," in *The nature and art of motion* (G. Kepes, ed.), Braziller, 1965. (Revised in H. Wallach, *On perception.* Quadrangle, 1976).

[30] J. Hochberg and E. McAlister, "A quantitative approach to figural goodness," *Journal of Experimental Psychology*, 46, 1953, 361–364.

[31] J. Hochberg and V. Brooks, "The psychophysics of form: Reversible perspective drawing of spatial objects," *American Journal of Psychology*, 73, 1960, 337–354.

[32] J. Cutting, "Coding theory adapted to gate perception," *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1981, 71–87.

[33] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, Mass.: MIT Press, 1979.

[34] D. D. Hoffman and B. E. Flinchbaugh, "The interpretation of biological motion," *Biological Cybernetics*, 42, 1982, 195–204.

[35] B. M. Bennett and D. D. Hoffman, "The computation of structure from fixed-axis motion: Nonrigid structures," *Biological Cybernetics*, 51, 1985, 293–300.

[36] J. Aloimonos, "Visual shape computation," *Proc. IEEE*, 76, 1988, 899–916.