Multi-View 3D Shape and Motion Recovery on the Spatio-Temporal Curve Manifold

Rodrigo L. Carceroni
Dept. of Computer Science
University of Rochester
Rochester, NY 14627 USA

Kiriakos N. Kutulakos
Depts. of Computer Science & Dermatology
University of Rochester
Rochester, NY 14627 USA

Abstract

In this paper we consider the problem of recovering the 3D motion and shape of an arbitrarily-moving, arbitrarily-shaped curve from multiple synchronized video streams acquired from distinct and known points in space. By studying the 3D motion and shape constraints provided by the input video streams, we show that (1) shape and motion recovery is equivalent to the problem of recovering the differential properties of the Spatio-Temporal Curve Manifold *that describes the curve's trace in space-time, and (2) a local analytical description of this manifold can be computed directly from the spatio-temporal volumes defined by the input video streams. Our experimental results suggest that this manifold-based approach to joint shape and motion estimation yields shape estimates of higher accuracy that those obtained from stereo alone, allows accurate recovery of 3D curve motion, and provides significant robustness against image noise and camera calibration errors.*

1. Introduction

There has been considerable interest in recovering the three-dimensional shape and motion of an unknown dynamic scene from sequences of images—e.g., work on optical flow [1], structure-from-motion [2], and 3D motion estimation [3]. One common characteristic of these approaches is the assumption that all images are acquired by a single camera. Unfortunately, because the scene is viewed from just a single viewpoint at a time, this assumption imposes strong constraints on the types of motion that can be recovered and on the scenes that can be analyzed. Existing work has therefore used a variety of additional assumptions to make 3D shape and motion estimation tractable (e.g., rigid [2,4,5], articulated [6], parametric [7], or isometric motion [8], known 3D shape [9], or known shape dynamics [3]).

In this paper we consider the problem of recovering the 3D shape and motion of a dynamic scene that is observed simultaneously by *multiple* cameras. In particular, we focus on the case where (1) the scene is composed of one or more smooth 3D curves that are moving rigidly or non-rigidly in space, (2) the curves' 3D shape and motion are completely unknown, and (3) the curves are viewed simultaneously by two or more synchronized cameras with known projection matrices. To study this problem, we introduce the notion of the *Spatio-Temporal Curve Manifold*, which is the trace of a moving 3D curve in space-time. We show that computing the shape and motion of an arbitrarily-moving 3D curve from multiple views is equivalent to the problem of recovering the local shape of this manifold at every point. Furthermore, we provide an efficient algorithm for recovering the manifold's shape directly from the spatio-temporal image volumes given as input. Our experimental results on both simulated and real scenes suggest that our manifold-based approach to joint shape and motion estimation yields shape estimates of higher accuracy that those obtained from stereo alone, leads to significant improvements in the accuracy of 3D motion measurements, and provides robustness against noise and camera calibration errors.

Little is currently known about how to recover the shape and motion of moving 3D curves in the multi-view case and about the constraints and ambiguities that this problem embodies. In a preliminary analysis, we showed that only two out of three components of the 3D motion of a curve point can be recovered from multi-view sequences, regardless of the number of cameras observing the scene [10]. Here we generalize that analysis by developing a differential-geometric framework in which the Spatio-Temporal Curve Manifold is used to capture implicitly all the shape and motion information about the scene that is recoverable from a multi-view sequence. Importantly, because this manifold can be estimated directly from the input images, the resulting algorithm does not rely on the accuracy of edge detection, linking, tangent estimation, and curve matching operations, which are sensitive to calibration errors and noise.

Our approach offers four key contributions over the existing state of the art. First, even though several systems have been developed for acquiring and processing multi-view sequences of dynamic scenes [11,12], existing recovery methods emphasize the mutual independence of the scene information available at different time instants. Approaches have therefore focused on (1) recovering 3D shape rather than motion, and (2) decomposing the 3D shape re-

covery problem for dynamic scenes into a collection of independent recovery problems, one for each instant in time. Unlike these approaches, our analysis leads to algorithms that recover 3D motion information from such sequences *and* exploit their spatio-temporal coherence to improve 3D shape estimates. Second, even though techniques for multi-view motion estimation have been proposed recently (e.g., [13,14]), these techniques rely on the availability of an *a priori* shape model to compute 3D motion. In contrast, our work does not rely on the availability of shape or motion models. Third, our work is specifically aimed at the analysis of 3D curve motion and, hence, it is closely related to previous work on curve-based stereo in static scenes [15,16] and on recovering shape from a single sequence of projections of a moving 3D curve [8,17–20]. Our analysis therefore generalizes these approaches to the case where the scene is dynamic, non-rigid, and simultaneously viewed by many cameras. Fourth, the Spatio-Temporal Curve Manifold represents the set of all 3D shape and motion solutions that are simultaneously consistent with the 2D shapes and image motions observed at every camera viewpoint. As such, our approach can be thought of as a novel application of the "intersection of constraints" paradigm [21,22] to the multi-view 3D shape and motion estimation problem.

2. The Spatio-Temporal Curve Manifold

Let γ be a regular curve [23] undergoing an unknown, smooth, and possibly non-rigid motion in space. We assume that γ is viewed simultaneously from $N \geq 2$ distinct viewpoints, $\mathbf{o}_i, i=1,\ldots,N$, by perspective cameras whose projection matrices are known. Our goal is to compute the 3D shape and motion of γ from its time-varying projections at the input viewpoints.

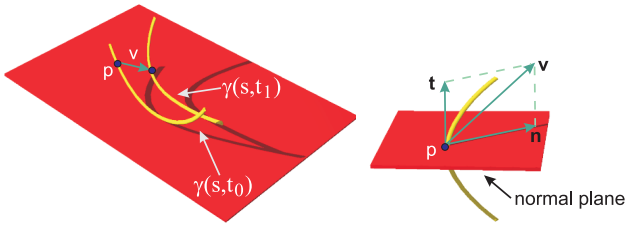
In particular, let $\gamma(s,t) : I \times (0,\infty) \rightarrow \Re^3$ be the unknown parameterization of γ that describes the 3D position and motion of every point in γ. Given a time $t_0 \in (0,\infty)$, $\gamma(s,t_0)$ is the curve's shape at t_0 (Figure 1a). Similarly, the curve $\gamma(s_0,t)$ is the spatio-temporal trajectory of an individual point on γ. We define the velocity, $\mathbf{v}$, of a point $\mathbf{p} = \gamma(s_0,t_0)$ to be equal to $\gamma_t(s_0,t_0)$, where γ_t is the partial derivative of γ with respect to t.

As the curve γ moves and deforms, it can be thought of as sweeping a surface in space. Mathematically, this surface is described by a two-dimensional manifold [24], Γ, that is embedded in space-time, i.e., $\Re^3 \times (0,\infty)$. We use the term *Spatio-Temporal Curve Manifold* to refer to this manifold; it generalizes to 3D curves the notion of the spatio-temporal surface, which has often been used to describe the temporal trajectory of 2D curves in the image plane [4,8,25,26].

The parameterization describing γ's shape and motion induces a parameterization of the manifold:

$$\Gamma : I \times (0,\infty) \rightarrow \Re^3 \times (0,\infty) \quad (1)$$

$$\Gamma(s,t) = [\gamma(s,t)\ t]. \quad (2)$$



(a) (b)

Figure 1. 3D motion geometry for a point $\mathbf{p}$ on γ.

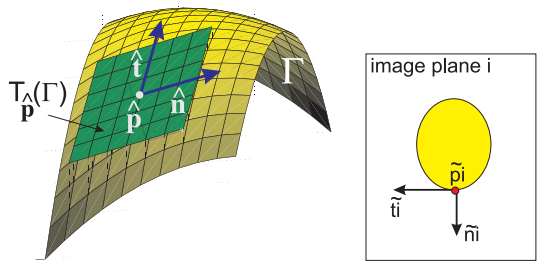
The partial derivatives, Γ_s and Γ_t, of this parameterization define $T_{\hat{\mathbf{p}}}(\Gamma)$, the tangent space of Γ at $\hat{\mathbf{p}} = [\mathbf{p}\ t_0]$. $T_{\hat{\mathbf{p}}}(\Gamma)$ is a plane embedded in space-time. Intuitively, it describes the curve's orientation at $\mathbf{p}$ as well as the way in which $\mathbf{p}$'s position changes through time.

The manifold Γ can always be expressed as the *envelope* of its tangent spaces [27]. As a result, the local shape of Γ at $\hat{\mathbf{p}}$ is completely described by $T_{\hat{\mathbf{p}}}(\Gamma)$ as well as the way in which $T_{\hat{\mathbf{p}}}(\Gamma)$ varies in $\hat{\mathbf{p}}$'s neighborhood. Here we exploit this observation in two ways. First, we show that the plane $T_{\hat{\mathbf{p}}}(\Gamma)$ captures all the information we can compute about the 3D shape and motion of γ at $\mathbf{p}$ from multiple views. Second, we formulate the problem of reconstructing the manifold Γ as the problem of (1) computing its tangent spaces directly from the image data, and (2) computing their envelope. We make this approach precise in the next section, where we relate the tangent space at $\hat{\mathbf{p}}$ to the curve's local 3D shape and motion at $\mathbf{p}$.

3. Differential Multi-View Constraints on 3D Shape & Motion

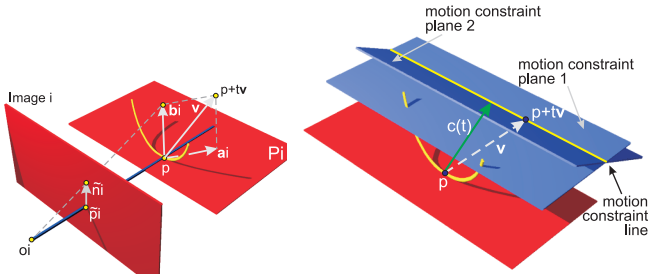
A basic step of our method for recovering 3D shape and motion from a multi-view image sequence is to establish a parameterization for the Spatio-Temporal Curve Manifold that captures all the 3D shape and motion constraints this sequence provides. When estimating this parameterization from images, the significance of this step becomes two-fold. First, it ensures that all the 3D shape and motion constraints provided by the input views are captured in the solution of the resulting estimation problem. Second, it ensures that this solution depends only on those geometric quantities of the curve's shape and motion that *can* be estimated from the input views. Both issues are important from a practical standpoint— the simultaneous satisfaction of all image constraints provides increased resistance to measurement errors, and parameterizations that are free of ambiguous quantities avoid instabilities in the estimation process.

More specifically, the velocity $\mathbf{v}$ of $\mathbf{p}$ can be decomposed into two components, $\mathbf{t}$ and $\mathbf{n}$, one along the curve's tangent and one on a plane that is normal to that tangent (Figure 1b). This decomposition leads to a parameterization of the tangent space $T_{\hat{\mathbf{p}}}(\Gamma)$ that is defined in terms of $\mathbf{p}$ and the



(a) (b)

Figure 2. (a) The $\mathbf{t}$–$\mathbf{n}$ parameterization of $T_{\hat{\mathbf{p}}}(\Gamma)$. (b) The projected displacement of $\mathbf{p}$ is decomposed into two orthogonal components, $\tilde{\mathbf{t}}_\mathbf{i}$ and $\tilde{\mathbf{n}}_\mathbf{i}$, along the tangent and normal of the projected curve at $\tilde{\mathbf{p}}_i$, respectively. Only $\tilde{\mathbf{n}}_\mathbf{i}$ is recoverable from the curve's time-varying projection [8].



(a) (b)

Figure 3. (a) Motion constraints from a single image [10]. The 3D component $\mathbf{b}_\mathbf{i}$ of $\mathbf{p}$'s displacement is completely determined by $\mathbf{p}$, its projection $\tilde{\mathbf{p}}_i$, the image displacement $\tilde{\mathbf{n}}_\mathbf{i}$, and the camera's projection matrix. (b) The point $\mathbf{p}+t\mathbf{v}$ is always contained in the Motion Constraint Line. The motion component $\mathbf{n}$ in Figure 1b is the limit of vector $c(t)$, which connects $\mathbf{p}$ to its closest point on the Motion Constraint Line.

two vectors $\frac{\mathbf{t}}{||\mathbf{t}||}$ and $\mathbf{n}$ (Figure 2a):

Definition 1 ($\mathbf{t}$–$\mathbf{n}$ Tangent Space Parameterization) *The tangent space $T_{\hat{\mathbf{p}}}(\Gamma)$ can described by the orthogonal parameterization*

$$\Pi(\lambda,\mu) = \hat{\mathbf{p}} + \lambda\,\hat{\mathbf{t}} + \mu\,\hat{\mathbf{n}}, \tag{3}$$

where $\lambda,\mu \in \Re$; $\hat{\mathbf{p}} = [\mathbf{p}\ t_0]$; $\hat{\mathbf{t}} = [\frac{\mathbf{t}}{||\mathbf{t}||}\ 0]$; and $\hat{\mathbf{n}} = [\mathbf{n}\ 1]$.

Corollary 1 (Local Manifold Parameterization) *A local parameterization of Γ can be derived from its first-order Taylor series expansion around $\hat{\mathbf{p}} = \Pi(0,0)$:*

$$\Gamma(s,t) = \Pi(s{-}s_0, t{-}t_0). \tag{4}$$

A key property of the $\mathbf{t}$–$\mathbf{n}$ parameterization is that $\mathbf{p}$, $\frac{\mathbf{t}}{||\mathbf{t}||}$, as well as $\mathbf{n}$ can be directly related to image measurements: $\mathbf{p}$ and $\frac{\mathbf{t}}{||\mathbf{t}||}$ depend on the instantaneous projection of the curve at the input viewpoints, while $\mathbf{n}$ depends on image velocities. We consider the geometry of these relationships below and show that these are the *only* relationships that can be recovered from images. In the following we assume that the image projections of $\mathbf{p}$ and $\mathbf{t}$ along viewpoint $\mathbf{o}_i$ are $\tilde{\mathbf{p}}_i$ and $\tilde{\mathbf{t}}_\mathbf{i}$, respectively, and that the normal image velocity at $\tilde{\mathbf{p}}_i$ is $\tilde{\mathbf{n}}_\mathbf{i}$ (Figure 2b).

3.1. Multi-View Constraints on 3D Shape

It is well known that two or more images of a static 3D curve taken from distinct viewpoints are sufficient to determine the curve's shape uniquely [15,28]. Intuitively, this is because the curve's projection at a given viewpoint, along with the camera's center of projection, define a ruled surface [23] that contains the curve; when many input views are available, the intersection of these ruled surfaces is the 3D curve itself. More precisely, the following observation relates the tangent $\mathbf{t}$ at $\mathbf{p}$ to the tangents at $\mathbf{p}$'s projection in the input views:

Observation 1 (Multi-View Shape Constraint) *If P_i is the plane defined by $\mathbf{o}_i$, $\tilde{\mathbf{p}}_i$, and $\frac{\tilde{\mathbf{t}}_\mathbf{i}}{||\tilde{\mathbf{t}}_\mathbf{i}||}$, $i = 1,\ldots,N$, and the planes $P_1,\ldots,P_N$ are not all identical, their intersection is a line tangent to γ at $\mathbf{p}$.*

3.2. Multi-View Constraints on 3D Motion

In order to establish the relationship between the velocity component $\mathbf{n}$ and the input views, we consider an alternative decomposition of $\mathbf{p}$'s velocity that is viewpoint-specific: $\mathbf{v}$ can be decomposed into a component $\mathbf{a}_i$ that lies on the plane P_i and a component $\mathbf{b}_i$ that is perpendicular to this plane (Figure 3a). This decomposition leads to three observations that relate the vector $\mathbf{n}$ to the input views. In particular, let $\mathbf{p}+t\mathbf{v}$ be a first-order Taylor series expansion of $\mathbf{p}$'s trajectory through time. Observation 2 tells us precisely what constraints on a point's 3D motion can be extracted from a single image and relates these constraints to the point's 3D position in space:

Observation 2 (Motion Constraint Plane) *[10] Every viewpoint $\mathbf{o}_i$ defines a unique* Motion Constraint Plane *that contains $\mathbf{p}+t\mathbf{v}$ and is determined by $\mathbf{p}$, $\tilde{\mathbf{p}}_i$, $\tilde{\mathbf{n}}_\mathbf{i}$, and the i-th camera's projection matrix. Moreover, this plane is the* only *constraint that can be imposed on $\mathbf{p}+t\mathbf{v}$ from the image at $\mathbf{o}_i$; it degenerates if and only if the optical ray through $\mathbf{p}$ contains the curve's tangent at $\mathbf{p}$.*

Intuitively, Observation 2 is based on the fact that the image at $\mathbf{o}_i$ provides no information about the component $\mathbf{a}_i$; the component $\mathbf{b}_i$, on the other hand, is completely determined by $\mathbf{p}$ and the curve's time-varying image at $\mathbf{o}_i$.

Observation 3 (Multi-View Motion Constraint Line) *[10] Given a time t, the intersection of any collection of non-degenerate and distinct Motion Constraint Planes is a unique*

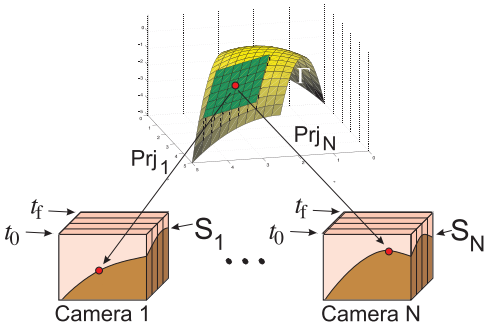


Figure 4. Manifold shape recovery strategy. Each viewpoint defines a spatio-temporal image volume depicting the moving curve. Our algorithm computes an analytical description of Γ's local shape directly from those N volumes.

line, called the Multi-View Motion Constraint Line at $\mathbf{p}$*, that is independent of the input viewpoints. This line contains the 3D position of* $\mathbf{p} + t\mathbf{v}$ *and is parallel to the curve's tangent at* $\mathbf{p}$.

Observation 3 tells us that the Motion Constraint Planes associated with a single point will in general form a pencil that constrains $\mathbf{p} + t\mathbf{v}$ to a single line (Figure 3b). This line captures all constraints on $\mathbf{p}$'s motion that can be derived from an arbitrary collection of views of a smooth curve and defines vector $\mathbf{n}$ uniquely:

Observation 4 (Multi-View Motion Constraint) *If* $\mathbf{c}(t)$ *is the vector that (1) starts at* $\mathbf{p}$*, (2) ends at the Motion Constraint Line, and (3) is perpendicular to it, the limit* $\lim_{t\to 0} \frac{\mathbf{c}(t)}{t}$ *is the vector* $\mathbf{n}$.

4. Manifold Reconstruction from Spatio-Temporal Image Volumes

The previous section suggests that we can estimate a curve's 3D shape and motion from multiple views by reconstructing the Spatio-Temporal Curve Manifold. Using this as a starting point, we describe a method for recovering a local analytic representation of this manifold at a point $\hat{\mathbf{p}}$. An important feature of the method is that it recovers the manifold's shape directly from the pixels in the multi-view sequence, i.e., without relying on intermediate edge- or curve-based representation stages to build the final reconstruction.

In order to achieve this, we treat each input video sequence as a 3D spatio-temporal volume of pixels and formulate manifold reconstruction as the problem of recovering a local analytical description of the manifold at $\hat{\mathbf{p}}$ from these volumes (Figure 4). The local shape at $\hat{\mathbf{p}}$ is recovered by first constructing an approximate, initial analytical description of the manifold, and then using this description to bootstrap a minimization procedure that refines it.

At the heart of this approach lies an image-based error functional that measures the consistency between the manifold's analytical description at $\hat{\mathbf{p}}$ and the pixels in the N spatio-temporal volumes. Intuitively, the functional compares pixels in these volumes to the values predicted by the current estimate of the manifold's shape. The functional is therefore fully specified by answering three questions: (1) how do we map points on the manifold to points in the spatio-temporal volume of each camera, (2) how do we identify which pixels in this volume are projections of the moving curve, and (3) how do we enforce consistency between the re-projected and actual images of the manifold?

To map an arbitrary 4D point $\hat{\mathbf{p}}$ in space-time to a unique 3D point in the i-th spatio-temporal volume, we use a trivial generalization of the perspective projection transformation—the three spatial coordinates of $\hat{\mathbf{p}}$ are mapped to two image coordinates using the known projection matrix of the i-th camera, while its temporal component remains unchanged:

$$\mathrm{Prj}_i(\hat{\mathbf{p}}) = [\tilde{\mathbf{p}}_i \; t].$$

Under ideal conditions, the image of Γ under this transformation should be identical to the spatio-temporal surface, $\mathcal{S}_i$, traced by the deforming projection of γ in the i-th camera. Our strategy is to recover an implicit analytical description of this surface from the pixel data, use the transformation Prj_i to "re-project" the manifold into each volume, and then define an error functional that compares the re-projected volume to the image measurements. More details about these steps are given below.

4.1. Recovering the Edge-Proximity Field

The first step in our method involves reconstructing from the spatio-temporal volume of camera i an analytical scalar field, $\mathcal{F}_i$, whose zero level set [29] is the spatio-temporal surface $\mathcal{S}_i$. The method does not depend on a specific definition of $\mathcal{F}_i$; the only requirements are that (1) $(\mathcal{F}_i)^2$ is differentiable in the neighborhood of $\mathcal{S}_i$, and (2) the zero set of $\mathcal{F}_i$ is identical to the projection of γ at the i-th viewpoint. In practice, we obtain the field by applying a Marr-Hildreth edge detector to each image in the spatio -temporal image volume. This edge detector convolves its input with a Laplacian kernel, producing a discrete scalar field whose zero-crossings are at the image intensity edges.

For every zero-crossing point, $[u_0\, v_0\, t_0]$, we use the field's discrete samples in the neighborhood of the point to compute a first-order analytical description of the underlying continuous field:

$$\begin{aligned}\mathcal{F}_i(u,v,t) \;=\; & \mathcal{F}_i(u_0,v_0,t_0) + \\ (u-u_0)\,\tfrac{\partial \mathcal{F}_i}{\partial u} \;+\; & (v-v_0)\,\tfrac{\partial \mathcal{F}_i}{\partial v} \;+\; (t-t_0)\,\tfrac{\partial \mathcal{F}_i}{\partial t}, \qquad (5)\end{aligned}$$

where the partial derivatives are all evaluated at $[u_0\, v_0\, t_0]$.

4.2. Parameterizing the Manifold's Tangent Space

We rely on the $\mathbf{t}$–$\mathbf{n}$ parameterization of Section 3 to define the manifold's tangent space in the neighborhood of a point $\hat{\mathbf{p}} \in \Gamma$. This parameterization has six degrees of freedom and is completely determined by (1) the parameters of the 3D line through $\mathbf{p}$ in the direction of $\mathbf{t}$, and (2) the parameters of the vector $\mathbf{n}$, which is perpendicular to this line. To specify the line's direction in space we use two angles, θ, ϕ; its position is specified by two coordinates, p_1, p_2, representing the point of intersection with a plane that is normal to the line and contains $\hat{\mathbf{p}}$. The vector $\mathbf{n}$, is specified by two coordinates, m_1, m_2, corresponding to a vector that is normal to the direction defined by θ and ϕ.

4.3. Defining the Re-Projection Error Functional

Given the $\mathbf{t}$–$\mathbf{n}$ parameterization of the tangent space at a point $\hat{\mathbf{p}} = \Pi(0,0)$, the projection transformation Prj_i of the i-th camera, and analytical descriptions of the edge proximity fields, $\mathcal{F}_i, i = 1, \ldots, N$, the re-projection error functional is defined as follows:

$$\mathcal{E}(\theta, \phi, p_1, p_2, m_1, m_2) = \int\int \sum_{i=1}^{N} \epsilon_i(s,t)\, ds\, dt, \qquad (6)$$

$$\text{where} \quad \epsilon_i(s,t) = \mathcal{F}_i^2\left(\mathrm{Prj}_i\left(\Pi(s,t)\right)\right), \qquad (7)$$

and the double integration is performed in a neighborhood of Γ around $\hat{\mathbf{p}}$ whose extent is specified *a priori*.

From a practical standpoint, this functional has two useful features. First, it can be evaluated very efficiently because a closed-form, analytical description of the functional is always available. Second, it does not impose restrictions on the number of input viewpoints and hence is particularly applicable to the analysis of shape and motion when large collections of cameras are observing the scene [11].

4.4. Reconstructing the Tangent Space

We use the re-projection error functional defined above to recover the parameters of a tangent space $T_{\hat{\mathbf{p}}}(\Gamma)$ of Γ that "passes near" an arbitrary 4D point in space-time. The Jacobian of $\mathcal{E}$ with respect to its parameter vector, $\mathbf{x} = [\theta\, \phi\, p_1\, p_2\, m_1\, m_2]$, is given by the chain rule:

$$\frac{\partial \mathcal{E}}{\partial \mathbf{x}} = \sum_{i=1}^{N} \frac{\partial \mathcal{E}}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\partial \mathcal{F}_i} \frac{\partial \mathcal{F}_i}{\partial \mathrm{Prj}_i} \frac{\partial \mathrm{Prj}_i}{\partial \Gamma} \frac{\partial \Gamma}{\partial \mathbf{x}}. \qquad (8)$$

Since an analytical description of $\epsilon_i(s,t)$ and its Jacobian are available, it is possible to use a variety of derivative-based optimization techniques to adjust $\mathbf{x}$ until $\mathcal{E}$ converges to a minimum value. We chose Levenberg-Marquardt [30] to perform this step.

5. Multi-View 3D Shape & Motion Recovery by Manifold Reconstruction

Section 4 leads directly to an algorithm for reconstructing the tangent spaces of the Spatio-Temporal Curve Manifold, for reconstructing the manifold itself, and for converting these reconstructions into 3D shape and motion estimates. The algorithm consists of the following steps:

Step 1: Apply the Marr-Hildreth edge detector to all input images and find all zero-crossings.

Step 2: Reconstruct the edge-proximity field in the neighborhood of every zero crossing, using Eq. (5).

Step 3: Choose a reference camera, $\mathbf{o}_1$, and repeat the following steps for every zero-crossing, $\tilde{\mathbf{p}}_1$, in the spatio-temporal volume of that camera:

- **a.** establish approximate stereo correspondences, $\tilde{\mathbf{p}}_2, \ldots, \tilde{\mathbf{p}}_N$, for $\tilde{\mathbf{p}}_1$ using the known epipolar geometry between cameras; use these correspondences to compute a 4D point $\hat{\mathbf{p}}$ near the manifold;
- **b.** compute approximate tangents and normal velocities, $\frac{\tilde{\mathbf{t}}_\mathbf{i}}{||\tilde{\mathbf{t}}_\mathbf{i}||}, \tilde{\mathbf{n}}_\mathbf{i}, \mathrm{i} = 1, \ldots, \mathrm{N}$, at the points $\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_N$; use these vectors to compute an initial estimate for the parameters of the tangent space of Γ near $\hat{\mathbf{p}}$;
- **c.** compute the tangent space parameters that minimize $\mathcal{E}(\theta, \phi, p_1, p_2, m_1, m_2)$.

Step 4: Reconstruct the manifold Γ as the envelope of all tangent spaces computed for zero crossings in the reference camera.

Step 5: Convert every reconstructed pair $(\hat{\mathbf{p}}, T_{\hat{\mathbf{p}}}(\Gamma))$ to a 3D curve point $\mathbf{p}$, a local tangent $\frac{\mathbf{t}}{||\mathbf{t}||}$, and a normal velocity vector, $\mathbf{n}$, using Eq. (3).

6. Experimental Results

6.1. Synthetic Scenes

In order to determine the applicability of our manifold-based approach, we performed experiments with several synthetic multi-view sequences. The purpose of our simulations was to test two hypotheses:

1. Joint estimation of 3D shape and motion yields motion estimates of significantly higher accuracy than those obtained by treating 3D motion computation as a post-processing step (i.e., after computing 3D shape from stereo).

2. Besides providing accurate 3D motion estimates, a joint estimation of 3D shape and motion improves the accuracy of 3D shape estimates as well.

More specifically, we used `OpenInventor` to generate three-view sequences consisting of six consecutive snapshots of a textured sphere that rotates and shrinks in front of a static background (Figure 5a). The distances between

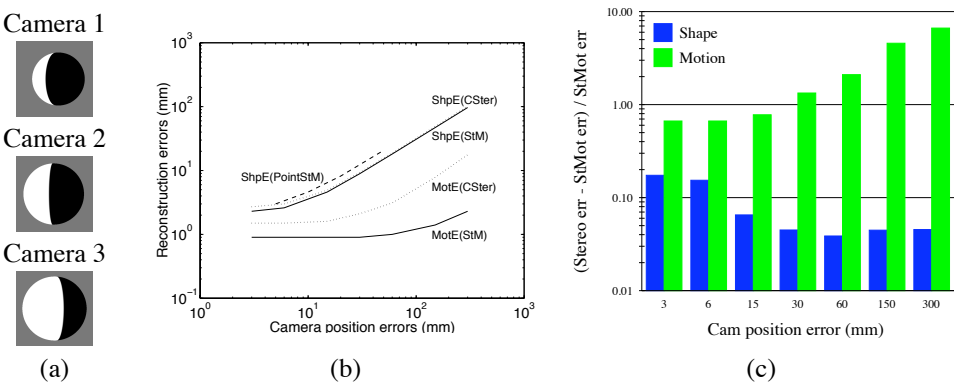


Figure 5. (a) Snapshots at the first time instant in a synthetic multi-view sequence used for our experiments. (b) Effects of camera calibration errors on the computed Shape (ShpE) and normal velocity estimates (MotE) obtained by our manifold reconstruction algorithm (StM), and by curve-based stereo (CSter). Camera calibration errors denote the norm of random displacement vectors added to the ground-truth position of the input cameras. A displacement parallel to the image plane of 30mm, which is 1% of the distance between the sphere's center and the cameras, roughly corresponds to a 5-pixel displacement in every input image. In addition to these simulated errors, the output of the Marr-Hildreth edge detector was corrupted with noise whose variance was 25% of the detector's maximum response on the input images. Shape reconstruction errors are measured as the RMS distance between points in the recovered curve and the ground-truth curve; motion errors are measured as the norm of the RMS difference between the recovered and ground-truth normal 3D motions of these points. All errors were averaged over 50 randomly-generated trials. For comparison purposes, the graph also plots reconstruction results for the point-based stereo-motion algorithm reported in [10] (PointStM). (c) Ratios between the errors obtained by our algorithm and those obtained by a curve-based stereo implementation.

the cameras and the center of the sphere were 3.6m, 3.2m and 3.0m, respectively, and the sphere's initial radius was 1m. The actual sequences were generated by (1) keeping constant the position of the input viewpoints, (2) simulating camera calibration errors by assigning a projection matrix to each camera that was derived by adding a random displacement to the camera's ground truth position, and (3) adding zero-mean Gaussian noise to the output of the Marr-Hildreth edge detector to simulate noise in the edge-proximity field (Section 4.1). This allowed us to evaluate the performance of our approach in the presence of both image intensity noise, which can affect localization of a curve's projections, and camera calibration errors.

To test our first hypothesis, we compared the manifold reconstruction algorithm of Section 5 to a curve-based stereo algorithm that differs from it only in Steps 3c, 4, and 5. Instead of employing these steps, we refined the initial shape estimates (i.e., 3D position and tangent at every point) using a curve-based stereo algorithm that fits a spline to these estimates. To compute motion estimates, we used a post-processing step in which planes normal to the reconstructed curve at each time t were intersected with the curve reconstructed at time $t+dt$. Both this technique as well as the algorithm of Section 5 were applied to approximately 250 points along the edge that separates the black and white regions on the synthetic sphere, at each of five consecutive time instants.

Figures 5b,c show how the 3D shape and motion reconstruction errors vary as a function of calibration errors. These results show that the estimates of the normal velocity, $\mathbf{n}$, reconstructed by our manifold-based algorithm are much more precise than those obtained by computing shape and motion independently. The improvements achieved through simultaneous estimation of shape and motion range from 67%, with calibration errors of 3mm, to 665%, with calibration errors of 300mm.

Our results suggest that the manifold-based approach also yields improvements in the accuracy of the reconstructed 3D shapes. For calibration errors of up to 6mm (which generate displacements of about a pixel in every input image), our algorithm's shape estimates are over 15% more accurate than those obtained through curve-based stereo. Improvements in shape estimates are reduced as calibration errors increase, but they remain close to 5% even for calibration errors as large as 300mm. Note that the manifold-based algorithm generates shape measurements of significantly higher accuracy than the point-based algorithm in [10], with relative improvements ranging from 20% to 37% for calibration errors of 5mm to 40mm, respectively. Importantly, this improvement was despite the fact that the errors in Figure 5b for the point-based algorithm occur for "ideal" input images with no noise in pixel intensities.

6.2. Discussion

Our algorithm's increasing advantage in estimating 3D motion in the presence of large calibration errors can be understood by considering how a curve's projection in multiple views constrains its 3D shape at a point $\mathbf{p}$ (Section 3.1). Ideally, the N planes defined by the image tangents at $\mathbf{p}$'s projection form a pencil whose common intersection is the curve's 3D tangent line. In the presence of calibration

errors, however, these planes will not have a single intersection and, due to the independent feature localization noise, it will only be possible to constrain the tangent's position to the interior of an uncertainty volume in the 3D scene space. If the positions of these tangents are computed independently at each time instant, the errors in the resulting normal velocity computations will be roughly proportional to the maximum extension of this uncertainty volume. On the other hand, if the tangent lines at times t and $t+dt$ are estimated jointly, the maximum displacement between them will be limited by the additional 3D motion constraints, regardless of the tangents' position within the uncertainty volume. Since the uncertainly volume is much larger than the inter-frame motions when the calibration errors are large and the sequences dense, a larger relative improvement in motion accuracy will be attained by an approach that explicitly enforces the (tighter) motion constraints.

Our results on 3D shape recovery suggest that a coupled estimation of shape and motion counteracts shape errors due to the localization of projected curve points. Specifically, since errors due to point localization are generally independent across images, the coupled estimation of shape and motion allows us to include images from multiple time instants in the shape estimation process, therefore providing additional independent constraints to counteract these errors. Large calibration errors, on the other hand, introduce a bias in shape reconstruction [31] that cannot be completely resolved from the additional constraints provided by image motion. As a result, the relative 3D reconstruction improvements obtained through joint shape/motion estimation is reduced as these errors increase.

6.3. A Real Scene

In order to evaluate the accuracy of our approach in practice, we used a 3-view image sequence displayed in Figure 6. The sequence was acquired using three calibrated Pulnix TMC-9700 progressive-scan color video cameras, each connected to a separate networked PentiumII PC equipped with a Matrox MeteorII real-time video capture card. All cameras were frame-synchronized using a VieCore video synchronization board. The cameras were positioned in a triangular configuration approximately 2m above the ground, approximately 50cm away from each other, and looking downward. A 15-frame (0.5sec) sequence was then captured while a letter-size sheet of paper was moved manually in the cameras' field of view so that one of the sheet's edges (the longer one, closer to the image center at Camera 1) remained in contact with the floor. We then used our manifold reconstruction algorithm to recover the shape and motion of this edge at each time instant.

Figure 7a shows $t = constant$ slices of the reconstructed spatio-temporal curve manifold. The manifold contains 4855 individual neighborhoods and took 2 minutes and 30 seconds to compute on an SGI Indigo2 workstation.

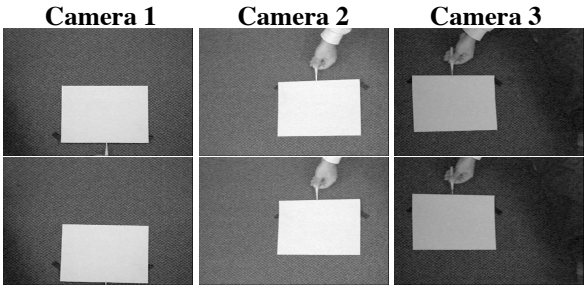


Figure 6. Snapshots at the first and last instants of a 15-frame, 3-view sequence used in our experiments.

Even visually, it is possible to verify that these slices form a planar surface, as expected. Using a Singular Value Decomposition, we determined the plane that most tightly fits all the reconstructed points; the average distance of the reconstructed points from this plane was only 0.5mm, confirming the almost-planar reconstruction. Figure 7b shows the computed normal motion field. This field is mostly composed of vectors with similar orientations and magnitudes, which is consistent with the fact that the paper sheet was roughly translated along a fixed direction, with a roughly fixed velocity. Note that the computed 3D shape and motion measurements are accurate and consistent throughout the sequence, despite the fact that all computations in our algorithm are purely local and assume no prior shape or motion models to drive the reconstruction process.

7. Concluding Remarks

A key remaining open question is how to recover the 3D shape and motion of dynamic scenes composed of arbitrarily-shaped and textured surfaces rather than curves. Toward this end, we are combining the geometric analysis in this paper with recent surface- and volume-based methods for multi-view shape recovery [22, 32] in order to investigate problem's underlying geometry and develop practical spatio-temporal reconstruction algorithms.

Acknowledgments The support of CAPES (Proc. 0591/95-2), of the National Science Foundation under Grant No. IRI-9875628, of Roche Laboratories, Inc., and of the Dermatology Foundation are gratefully acknowledged. We would also like to thank Craig Harman for his invaluable help in designing, building, and calibrating our experimental multi-camera system.

References

[1] M. J. Black, "Explaining optical flow events with parameterized spatio-temporal models," in *Proc. CVPR*, pp. 326-332, 1999.

[2] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *IJCV*, v. 9, n. 2, pp. 137-154, 1992.

[3] D. DeCarlo and D. Metaxas, "Deformable model-based shape and motion analysis from images using motion residual error," in *Proc. 6th ICCV*, pp. 113-119, 1998.

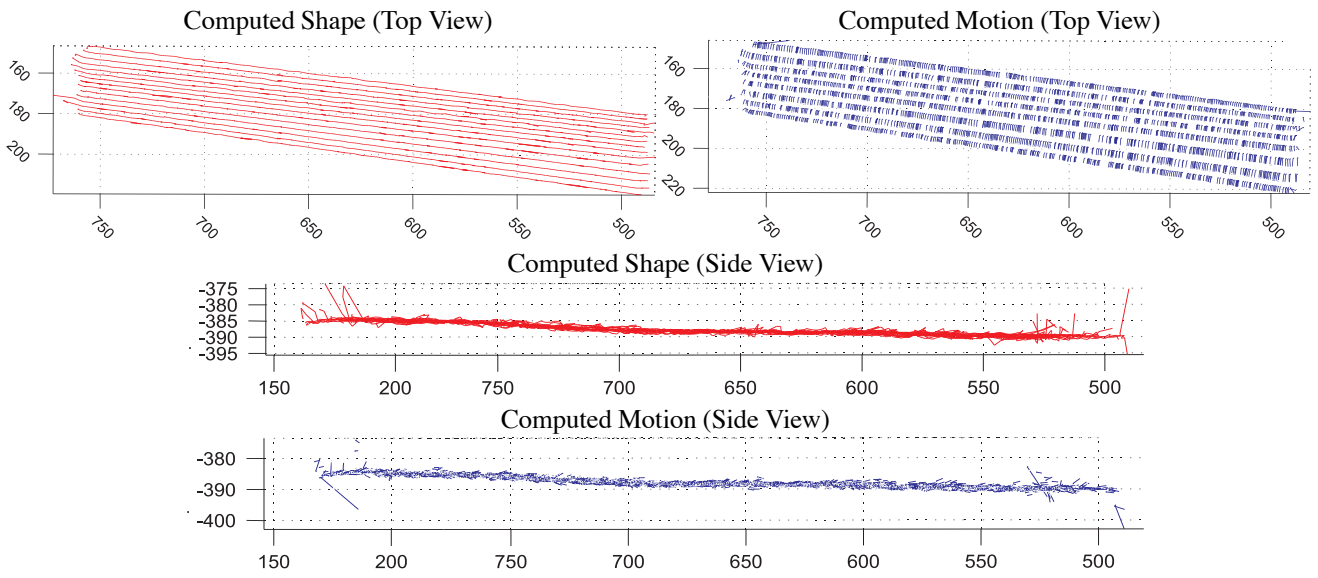


Figure 7. Views of the reconstructed 3D shapes and motions for the sequence in Figure 6. Connected curves in the upper-left figure correspond to fixed-time slices of the reconstructed manifold. The reconstructed normal motion field (i.e., the vectors $\mathbf{n}$), is shown only for the odd instants of the sequence. All scales are in millimeters.

[4] H. H. Baker, "Generalizing epipolar–plane image analysis on the spatiotemporal surface," *IJCV*, v. 3, pp. 33-49, 1989.

[5] P. K. Ho and R. Chung, "Stereo-motion that complements stereo and motion analyses," in *Proc. CVPR*, pp. 213-218, 1997.

[6] D. G. Lowe, "Fitting parameterized three–dimensional models to images," *IEEE T-PAMI*, v. 13, n. 5, pp. 441-450, 1991.

[7] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet, "Learning parameterized models of image motion," in *Proc. CVPR*, pp. 561–567, 1997.

[8] O. Faugeras and T. Papadopoulo, "A theory of the motion fields of curves," *IJCV*, v. 10, n. 2, pp. 125-156, 1993.

[9] H. Araújo, R. L. Carceroni, and C. M. Brown, "A fully projective formulation to improve the accuracy of Lowe's pose–estimation algorithm," *CVIU*, v. 70, n. 2, pp. 227-238, 1998.

[10] R. L. Carceroni and K. N. Kutulakos, "Toward recovering shape and motion of 3D curves from multi-view image sequences," in *Proc. CVPR*, pp. 192-197, 1999.

[11] P. J. Narayanan, P. W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," in *Proc. 6th ICCV*, pp. 3-10, 1998.

[12] S. Moezzi, "Immersive telepresence," *IEEE Multimedia*, v. 4, n. 1, pp. 17-26, 1997.

[13] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Proc. CVPR*, pp. 8-15, 1998.

[14] B. Guenter, C. Grimm, H. Malvar, and D. Wood, "Making faces," in *Proc. SIGGRAPH*, 55–66, 1998.

[15] B. Bascle and R. Deriche, "Stereo matching, reconstruction and refinement of 3D curves using deformable contours," in *Proc. 4th ICCV*, pp. 421-430, 1993.

[16] J. Porrill and S. Pollard, "Curve matching and stereo calibration," *IVC*, v. 9, n. 1, pp. 45–50, 1991.

[17] R. Cipolla and A. Blake, "Surface shape from the deformation of apparent contours," *IJCV*, v. 9, n. 2, pp. 83-112, 1992.

[18] R. Cipolla and A. Zisserman, "Qualitative surface shape from deformation of image curves," *IJCV*, v. 8, n. 1, pp. 53–69, 1992.

[19] K. N. Kutulakos and C. R. Dyer, "Global surface reconstruction by purposive control of observer motion," *Artif. Intell.*, v. 78, no. 1-2, pp. 147-177, 1995.

[20] K. Astrom, R. Cipolla, and P. J. Giblin, "Generalized epipolar constraints," in *Proc. 4th ECCV*, pp. 97-108, 1996.

[21] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome, "The analysis of moving visual patterns," *Exper. Brain Res.*, v. 11, pp. 117-152, 1986.

[22] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving." In these proceedings.

[23] M. P. doCarmo, *Differential Geometry of Curves and Surfaces*. Englewood Cliffs, NJ: Prentice-Hall Inc., 1976.

[24] M. P. doCarmo, *Riemannian Geometry*. Boston, MA: Birkauser, 1992.

[25] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *IJCV*, v. 1, pp. 7-55, 1987.

[26] T. Papadopoulo and O. Faugeras, "Motion field of curves: Applications," in *Proc. 3rd ECCV*, pp. 71-82, 1994.

[27] J. W. Bruce and P. J. Giblin, *Curves and Singularities*. Cambridge University Press, 2nd ed., 1992.

[28] O. Faugeras and L. Robert, "What can two images tell us about a third one?," in *Proc. 3rd ECCV*, pp. 485-492, 1994.

[29] J. A. Sethian, *Level Set Methods*. Cambridge University Press, 1996.

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

[31] G. Baratoff and Y. Aloimonos, "Changes in surface convexity and topology caused by distortions of stereoscopic visual space," in *Proc. 5th ECCV*, pp. 226-240, 1998.

[32] O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," in *Proc. 5th ECCV*, pp. 379-393, 1998.