International Journal of Computer Vision 49(2/3), 175–214, 2002

Multi-View Scene Capture by Surfel Sampling: From Video Streams to Non-Rigid 3D Motion, Shape and Reflectance*

RODRIGO L. CARCERONI
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, CEP 31270-010, Brazil
carceron@dcc.ufmg.br

KIRIAKOS N. KUTULAKOS
Department of Computer Science, University of Toronto, Toronto, ON M5S3H5, Canada
kyros@cs.toronto.edu

Received March 21, 2001; Revised November 9, 2001; Accepted January 15, 2002

Abstract. In this paper we study the problem of recovering the 3D shape, reflectance, and non-rigid motion properties of a dynamic 3D scene. Because these properties are completely unknown and because the scene's shape and motion may be non-smooth, our approach uses multiple views to build a piecewise-continuous geometric and radiometric representation of the scene's trace in space-time. A basic primitive of this representation is the *dynamic surfel*, which (1) encodes the instantaneous local shape, reflectance, and motion of a small and bounded region in the scene, and (2) enables accurate prediction of the region's dynamic appearance under known illumination conditions. We show that complete surfel-based reconstructions can be created by repeatedly applying an algorithm called Surfel Sampling that combines sampling and parameter estimation to fit a single surfel to a small, bounded region of space-time. Experimental results with the Phong reflectance model and complex real scenes (clothing, shiny objects, skin) illustrate our method's ability to explain pixels and pixel variations in terms of their underlying causes—shape, reflectance, motion, illumination, and visibility.

Keywords: stereoscopic vision, 3D reconstruction, multiple-view geometry, multi-view stereo, space carving, motion analysis, multi-view motion estimation, direct estimation methods, image warping, deformation analysis, 3D motion capture, reflectance modeling, illumination modeling, Phong reflectance model

1. Introduction

In this paper we consider the problem of *Multi-View Scene Capture*—using multiple cameras to simultaneously recover the shape, reflectance and non-rigid motion of an unknown scene that evolves through time in a completely unknown way. While many techniques exist for recovering one of these properties when the rest of them are known (e.g., capturing the 3D motion of articulated (Drummond and Cipolla, 2000; Deutscher et al., 2000; Sidenbladh et al., 2000; Delamare and Faugeras, 1999; Yacoob and Davis, 2000), or deformable scenes (DeCarlo and Metaxas, 2000; Guenter et al., 1998; Zhou and Kambhamettu, 2000); reconstructing static Lambertian scenes (Kutulakos and Seitz, 1999; Szeliski, 1999; Brodsky et al., 1999); and recovering the reflectance of static scenes with known shape (Sato et al., 1997; Wood et al., 2000)), our focus here is on the general case. In particular, how can we capture 3D scenes whose appearance depends on

*This research was conducted while the authors were with the Departments of Computer Science and Dermatology at the University of Rochester, Rochester, NY, USA.

time-varying interactions between shape, reflectance, illumination, and motion? Answering this question would go a long way toward recovering many common real-world scenes that are beyond the current state of the art, including highly-deformable and geometrically-complex surfaces whose shape, motion, self-occlusions and self-shadows change through time (e.g., clothing (Baraff and Witkin (1998))), non-Lambertian surfaces with complex shape and deformation properties (e.g., mm-scale dynamic representations of the human body), and static or moving 3D objects with specular surfaces (Lin and Lee, 1999, 2000; Lu et al., 1999; Sato and Ikeuchi, 1994; Nayar et al., 1993; Torrance and Sparrow, 1967).

We argue that general solutions to the scene capture problem must ultimately satisfy four criteria:

- *Generality*: Computations should rely as little as possible on the scene's true motion, shape and reflectance.
- *Physical consistency*: Computations should consistently explain all pixels and pixel variations in terms of their *physical causes*, i.e., the 3D position, orientation, 3D velocity, visibility, and illumination of individual scene points (which can change dramatically), and their reflectance (which usually does not).
- *Reconstructibility conditions*: It should be possible to state the conditions under which the reconstruction process is valid and/or breaks down.
- *Spatial and temporal coherence*: Real scenes rarely consist of isolated and independently-moving points and therefore this constraint should be integrated with computations.

As a first step in this direction, we present a novel mathematical framework whose goal is to recover a piecewise-continuous geometric and radiometric representation of the space-time trace of an unknown scene. The representation's basic primitive is the *dynamic surfel* (*surf*ace *el*ement), a high-degree-of-freedom description of shape, reflectance and motion in a small, bounded 4D neighborhood of space-time. Dynamic surfels encode the instantaneous position, orientation, curvature, reflectance, and motion of a small region in the scene, and hence enable accurate prediction of its static and dynamic appearance under known illumination conditions. Using this representation as a starting point, we show that scene capture can be achieved by formulating and solving a collection of spatiotemporally-distributed optimization problems, each of which attempts to recover a single dynamic surfel that approximates the scene's shape, reflectance and motion in a specific space-time neighborhood.

At the heart of our approach lie two key observations. The first observation is that when an opaque scene is viewed and illuminated in a known way and its reflectance is defined parametrically, it is possible to determine the consistency of a surfel with the input views regardless of the complexity of the scene's shape or its reflectance function, as long as the inter-reflections can be ignored. Using this observation as a starting point, we reduce instantaneous shape recovery to the problem of performing a sequence of *space queries*. Each query determines whether any scene points exist inside a specific bounded neighborhood of 3D space and, if they do, it computes the globally-optimal surfel fit, i.e., the surfel that best predicts the colors at the points' projections. We show that every query defines a global optimization problem in the space of all surfel descriptions and that it is possible to quantify precisely the conditions under which the globally-optimal solution is consistent with the appearance of the true scene. Importantly, we show that we can efficiently search surfel space for this solution with an algorithm called *Surfel Sampling*. This algorithm integrates explicit sampling of surfel space with a sequence of linear and non-linear parameter estimation stages to find the optimal surfel fit. Moreover, by combining Surfel Sampling with a global method that resolves camera and light-source occlusions, we can capture 3D scenes despite dramatic changes in the visibility and appearance of scene points.

The second observation is that 3D motion recovery becomes considerably simplified when the scene's instantaneous 3D shape and reflectance properties have already been estimated. Starting from the principle that *reflectance* is the only scene property that remains constant, we show that we can (1) recover 3D motion descriptions without making any assumptions about the motion of scene points, (2) estimate 3D motion even in the presence of moving specularities, (3) incorporate spatio-temporal coherence into motion computations for improved stability, (4) assign a dense and non-rigid instantaneous motion field to every surfel by solving a direct linear estimation problem that depends only on pixel intensities and generalizes existing direct methods (Zelnik-Manor and Irani, 2000; Irani, 1999), and (5) improve shape and reflectance estimates by incorporating dynamic constraints into the surfel estimation process. Experimental results with real scenes that deform in very complex ways (clothing, skin) and the Phong reflectance model (Watt, 2000) illustrate

the method's ability to recover coherent 3D motion, shape and reflectance estimates from multiple views.

Little is currently known about how to recover simultaneously such estimates for unknown, geometrically-complex and deforming scenes. Since our goal is to recover these estimates simultaneously, our work is closely related to approaches in both stereo and motion estimation. While recent scene-space stereo methods can successfully recover the shape of complex 3D scenes by resolving multi-view occlusion relationships, they rely on discrete shape representations (e.g., lines (Collins, 1996), voxels (Kutulakos and Seitz, 2000; Seitz and Dyer, 1999; Kutulakos, 2000; Narayanan et al., 1998; Roy and Cox, 1998; Chen and Medioni, 1999), and layers (Szeliski, 1999; Szeliski and Golland, 1998; Szeliski et al., 2000)) that cannot model surface orientation explicitly. This has limited their applicability to Lambertian scenes with static illumination, where the dependencies between surface orientation, scene illumination, and scene appearance can be ignored. Moreover, their built-in emphasis on pixel-wise correspondence metrics (Kutulakos and Seitz, 1999) makes it difficult to incorporate spatial coherence constraints which can reduce sensitivity to noise and can lead to physically-plausible reconstructions (Snow et al., 2000). Even though mesh, particle and level-set stereo methods (DeCarlo and Metaxas, 2000; Fua and Leclerc, 1995; Fua, 1997, 1999; Samaras and Metaxas, 1998; Jin et al., 2000; Faugeras and Keriven, 1998) can, in principle, model surface orientation (Samaras and Metaxas, 1998; Jin et al., 2000), their use of a single functional to assess the consistency of a complete shape makes it difficult to study how reflectance and illumination computations at one location of a shape will affect computations elsewhere. Even more importantly, no formal study has been undertaken to establish the reconstructibility conditions of these techniques, i.e., the conditions under which the computed reconstructions coincide with the scene's true shape.

Unlike existing stereo methods, our surfel-based representation and our space-query formalism are spatially localized. This allows us to define a tractable optimization problem with well-defined reconstructibility conditions and an algorithm that has predictable performance. Moreover, because surface orientation is explicitly represented, our approach can handle scenes with non-Lambertian reflectance and multiple sources of illumination. Hence, our surfel-based formalism can be thought of as striking a balance between the need to reason about orientation explicitly during the reconstruction process, the desire to incorporate the spatial coherence constraints found in many binocular stereo techniques (Ohta and Kanade, 1985; Belhumeur, 1996; Silva and Santos-Victor, 2000) while preserving the occlusion-resolution properties of scene-space methods, and the desire to ensure the tractability of the shape estimation problem by keeping the representation of distant scene points separate.

In the context of motion estimation, single-view methods have relied on known models of 3D shape (Drummond and Cipolla, 2000; Deutscher et al., 2000; Sidenbladh et al., 2000; Delamare and Faugeras, 1999; Guenter et al., 1998; Lowe, 1991; Bregler and Malik, 1998; DeCarlo and Metaxas, 1998) or 3D motion (Sidenbladh et al., 2000; Zelnik-Manor and Irani, 2000; Irani, 1999; Tomasi and Kanade, 1992; Bregler, 2000; Ben-Ezra et al., 2000; Avidan and Shashua, 2000; Bérériat et al., 2000; Zhou et al., 2000; Papin et al., 2000) to make motion estimation a well-posed problem, or have focused on improving the robustness (Anandan, 1989; Black et al., 2000; Ye and Haralick, 2000; Smith et al., 2000) and physical validity (Fleet et al., 2000; Haussecker and Fleet, 2000; Negahdaripour, 1998; Gaucher and Medioni, 1999) of 2D motion estimation. Unfortunately, the use of known 3D shape models limits the types of scenes that can be reconstructed, while the ill-posed nature of single-view 3D motion estimation makes it difficult to account for image variations in a way that is consistent with a scene's true 3D geometry (Haussecker and Fleet, 2000). Even though a small number of multi-view methods has been proposed for estimating 3D motion, their reliance on potentially-noisy pointwise flow calculations (Vedula et al., 1999, 2000) and on the brightness constancy assumption (Vedula et al., 1999, 2000; Zhang and Kambhamettu, 2000; Tzovaras and Grammalidis, 1997) restricts them to slowly-moving Lambertian scenes, where the effects of shading and shadows on scene appearance is negligible.

Our approach extends existing methods to 3D motion capture in six ways. First, by relying on a spatially-distributed surfel representation to compute unconstrained 3D motion, our approach allows us to reconstruct 3D motion fields that are beyond the capabilities of existing model-based methods. Second, because these fields are dense and are extracted from video alone, they allow us to analyze the motion of scenes, such as clothing, whose motion cannot be adequately represented by a small number of feature points (as in the case of articulated figures) and whose mechanical

properties would change if they were instrumented to facilitate motion capture (e.g., with optical or magnetic trackers). Third, by relying on multiple views to recover 3D motion, our approach leads to a well-posed and solvable estimation problem. Fourth, because our approach recovers a parametric 3D motion field independently for each dynamic surfel, it can be thought of as generalizing existing methods that rely on 2D layers (Wang and Adelson, 1993) and parameterized 2D flow models (Ye and Haralick, 2000; Black and Anandan, 1996; Black, 1999) to exploit spatio-temporal coherence and reduce sensitivity to noise. Fifth, by relying on reflectance- instead of brightness-constancy, our formulation allows us to recover the motion of rapidly-moving Lambertian scenes as well as scenes with non-Lambertian reflectance properties. Sixth, by integrating 3D shape, reflectance, and motion computations into a single optimization framework, our approach ensures that all multi-view and multi-frame constraints contribute simultaneously to the estimation of the scene's static and dynamic properties.

The rest of the paper is structured as follows. Section 2 describes our image formation model and leads to a pair of *picture invariants*, which allow us to extend the common brightness constancy assumption to moving Lambertian or non-Lambertian scenes. Sections 3 and 4 then describe our dynamic surfel representation and define the notion of *surfel photo-consistency*, which allows us to quantify the consistency between a surfel-based scene description and the multi-view image streams that are given as input. These sections lead to a key reconstructibility result, the Surfel Approximation Theorem, that characterizes the conditions under which surfel photo-consistency can be established and that motivates our surfel-based reconstruction approach. Using this theorem as a starting point, Section 5 introduces our space query framework and describes the Surfel Sampling Algorithm for computing scene shape and reflectance. This framework is extended in Section 6 through the development of a non-linear method for surfel-based 3D motion estimation. Section 7 then summarizes our algorithm for global 3D shape, reflectance and motion recovery, and Section 8 presents experimental results on a variety of complex real scenes.

2. Picture Invariants

Consider a 3D scene undergoing unknown and potentially non-rigid motion in space. We assume that the scene is viewed under perspective projection from $N \geq 2$ known positions $\mathbf{c}_1, \ldots, \mathbf{c}_N$ and is illuminated by L known point light sources $\mathbf{l}_1, \ldots, \mathbf{l}_L$. Our goal is to compute the scene's 3D shape and motion from its time-varying projections in the N input views. To achieve this, we define a set of *picture invariants* that allow us to relate the instantaneous colors and intensities of pixels in one view of the scene to those observed in other views and time instants. We obtain these invariants by first modeling the way that the scene's geometry, motion, and reflectance properties determine its appearance (Fig. 1).

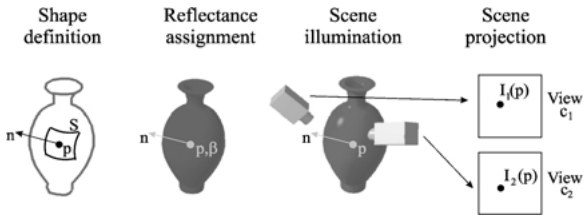

Figure 1. Steps defining the scene's image formation model.

2.1. Static and Dynamic Scene Geometry

We assume that the scene's instantaneous shape is described by a collection of regular surfaces whose closure bounds a possibly-disconnected volume in $\Re^3$ (do Carmo, 1976). When a regular surface S of the scene moves or deforms in space, the spatio-temporal evolution of its points can be described by a three-parameter function $\hat{\mathbf{x}}$ that encodes the points' 3D position and velocity (Fig. 2(a)). Let $\mathbf{x} : (U \subset \Re^2) \rightarrow \Re^3$ be an orthonormal parameterization of the surface S at time $t = t_0$. The evolution of S through time can then be

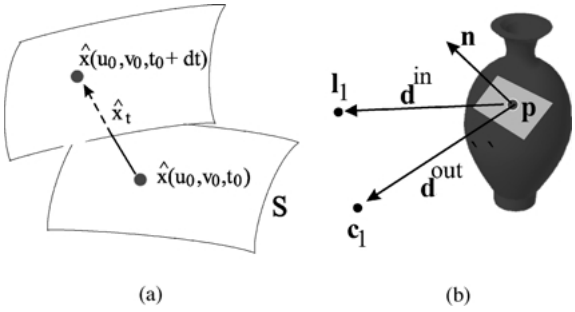

Figure 2. (a) Dynamic scene geometry. (b) Scene reflectance geometry.

described by a function $\hat{\mathbf{x}}(u, v, t) : (U \times T) \rightarrow \Re^3$ that simply extends $\mathbf{x}$ into the time dimension. Given a time instant $t_1 \in T$ potentially distinct from t_0, the function $\hat{\mathbf{x}}(u, v, t_1)$ is the parameterization of S's instantaneous shape at time t_1. Similarly, given a moving surface point $\mathbf{p}$ whose instantaneous 3D position at t_0 is $\mathbf{x}(u_0, v_0)$, the function $\hat{\mathbf{x}}(u_0, v_0, t)$ defines $\mathbf{p}$'s space-time trace. In the following, we focus our attention on dynamic scenes for which the function $\hat{\mathbf{x}}$ is smooth everywhere. This allows us to assign a well-defined velocity, $\hat{\mathbf{x}}_t$, to all points on S for all time instants in T.[1]

2.2. Image Formation Model

The color and intensity at the projection of a visible scene point depend on the scene's reflectance properties and the point's incident illumination. We rely on the Phong model (Watt, 2000) to represent the reflectance properties of specular scenes, one of the mathematically- and computationally-simplest parametric reflectance models available. The model implicitly specifies the scene's bi-directional reflectance distribution function (BRDF) (Lin and Lee, 1999; Torrance and Sparrow, 1967; Wolff et al., 1998; Oren and Nayar, 1997; Koenderink et al., 1999; Lafortune et al., 1997; Cook and Torrance, 1981), i.e., the ratio of outgoing radiance along a direction $\mathbf{d}^{\text{out}}$ to incoming irradiance along a direction $\mathbf{d}^{\text{in}}$ as a function of the local surface geometry (Fig. 2(b)).

Let $\mathbf{p} = \mathbf{x}(u, v)$ be a scene point with normal $\mathbf{n} = \mathbf{x}_u \wedge \mathbf{x}_v$. We define the reflectance model of $\mathbf{p}$ to be the sum of two components, a diffuse reflection component, β^{D}, and a specular reflection component, β^{S}, that follows a cosine-lobe law:

$$\beta(\mathbf{p}, \mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}}, \lambda) \stackrel{\text{def}}{=} \beta^{\text{D}}(\mathbf{p}, \mathbf{n}, \mathbf{d}^{\text{in}}, \lambda) + \beta^{\text{S}}(\mathbf{p}, \mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}}) \quad (1)$$

$$\beta^{\text{D}}(\mathbf{p}, \mathbf{n}, \mathbf{d}^{\text{in}}, \lambda) \stackrel{\text{def}}{=} \rho(u, v, \lambda) C^{\text{D}}(\mathbf{n}, \mathbf{d}^{\text{in}}) \quad (2)$$

$$\beta^{\text{S}}(\mathbf{p}, \mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}}) \stackrel{\text{def}}{=} f(u, v)[C^{\text{S}}(\mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}})]^{k(u,v)} \quad (3)$$

where λ corresponds to a color band (red, green, or blue);[2] $\rho(u, v, \lambda)$ defines the surface albedo at $\mathbf{p}$; $f(u, v)$ and $k(u, v)$ are the coefficients of the specular model; $C^{\text{D}}(\mathbf{n}, \mathbf{d}^{\text{in}})$ is the cosine of the angle between the normal $\mathbf{n}$ and $\mathbf{d}^{\text{in}}$; and $C^{\text{S}}(\mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}})$ is the cosine of the angle between $\mathbf{d}^{\text{out}}$ and the reflection of $\mathbf{d}^{\text{in}}$ about the normal $\mathbf{n}$. We assume that the functions $f(u, v)$ and $k(u, v)$ are arbitrary piecewise-smooth functions and that the albedo function, $\rho(u, v)$, has a finite power spectrum but is arbitrary in all other respects.

Knowledge of the visibility, incident illumination, 3D position, orientation and reflectance properties of every scene point $\mathbf{p}$ is sufficient to reproduce images of the scene for any camera and light source position. We call $\mathbf{p}$ visible from a viewpoint $\mathbf{c}_i$ if and only if the open line segment $\mathbf{pc}_i$ does not intersect the scene volume. The scene's reflectance model tells us that the pixel color at the i-th projection of any visible point is a sum of two colors, a *diffuse color* $I^{\text{D}}(\mathbf{p})$ and a *specular color* $I_i^{\text{S}}(\mathbf{p})$:

$$I_i(\mathbf{p}) \stackrel{\text{def}}{=} I^{\text{D}}(\mathbf{p}) + I_i^{\text{S}}(\mathbf{p}) \quad (4)$$

$$I^{\text{D}}(\mathbf{p}) \stackrel{\text{def}}{=} \sum_{l=1}^{L} \beta^{\text{D}}(\mathbf{p}, \mathbf{n}, \mathbf{l}_l - \mathbf{p}) \mathcal{L}_l(\mathbf{p}) \quad (5)$$

$$I_i^{\text{S}}(\mathbf{p}) \stackrel{\text{def}}{=} \sum_{l=1}^{L} \beta^{\text{S}}(\mathbf{p}, \mathbf{n}, \mathbf{c}_i - \mathbf{p}, \mathbf{l}_l - \mathbf{p}) \mathcal{L}_l(\mathbf{p}) \quad (6)$$

where $\mathcal{L}_l(\mathbf{p})$ measures the irradiance due to the l-th light source as a function of point position and light source color. In our model, we assume that $\mathcal{L}_l(\mathbf{p})$ is zero if $\mathbf{p}$ is in shadow from $\mathbf{l}_l$ (i.e., $\mathbf{p}$ is not visible from position $\mathbf{l}_l$) and is a known function of $\|\mathbf{l}_l - \mathbf{p}\|$ otherwise. Hence, Eqs. (5) and (6) tell us that the diffuse and specular colors at the projection of a point $\mathbf{p}$ are obtained by summing the diffuse and specular contributions, respectively, of every light source that directly illuminates point $\mathbf{p}$. This model assumes that (1) pixel intensities measure scene radiance directly, (2) the scene is opaque, and (3) inter-reflections can be ignored.[3]

The image formation model defined above has four important properties. First, it allows us to capture viewpoint-dependent pixel color variations that are common in scenes with non-diffuse reflectance properties. Second, it allows us to represent the appearance of scenes under very general viewing and illumination conditions—the scene may self-occlude, may contain multiple point light sources, may cast shadows onto itself from one or more light sources, and may exhibit multiple specular highlights due to potentially-distinct light sources. Third, it requires only a small number of parameters—the albedo ρ, the factor f that defines the weight of the specular component, and the specular exponent k—to fully specify the appearance of a scene point under known illumination. This is especially important from a computational point of view, since these parameters must ultimately be estimated from the input

photographs using non-linear estimation.[4] Fourth, it models the color and intensity of every pixel as the sum of two colors, one of which depends on viewpoint and one of which does not. This leads directly to our notion of a *picture invariant*, described below, which aims to eliminate the viewpoint- and time-dependent components of a point's color in the input views.

2.3. Static and Dynamic Picture Invariants

Even though our image formation model suggests that the color and intensity at the projection of a scene point can vary from one viewpoint to another or one time instant to the next, these variations cannot be arbitrary. Here we use this observation to assign to every pixel in the input view one static and one dynamic *picture invariant*, i.e., a color that is viewpoint- and time-independent, respectively.

In particular, let $\mathbf{p} = \mathbf{x}(u_0, v_0)$ be a scene point at time $t = t_0$ and let $I_i(\mathbf{p})$ be the pixel color at $\mathbf{p}$'s projection along a viewpoint $\mathbf{c}_i$ from which $\mathbf{p}$ is visible. Observation 1 tells that as long as we can compute the contribution of specular reflectance to $\mathbf{p}$'s appearance, we can define a color $I_i^{\mathrm{Dinv}}(\mathbf{p})$ that does not depend on the camera's viewpoint:

Observation 1 (Static Picture Invariant).

$$I_i^{\mathrm{Dinv}}(\mathbf{p}) \stackrel{\mathrm{def}}{=} I_i(\mathbf{p}) - I_i^S(\mathbf{p}) = I^{\mathrm{D}}(\mathbf{p}), \tag{7}$$

where $I_i^S(\mathbf{p})$ is given by Eq. (6).

Intuitively, the Static Picture Invariant $I_i^{\mathrm{Dinv}}(\mathbf{p})$ estimates the contribution of diffuse reflectance to $\mathbf{p}$'s appearance, which is a viewpoint-independent quantity. Since this invariant is defined for every pixel in an input view, it can be thought of as defining, for every input image I_i, a new image I_i^{Dinv} whose pixel colors and intensities do not depend on the viewpoint from which the original image was taken.[5]

Our Static Picture Invariant can also be thought of as a generalization of the commonly-used *brightness constancy constraint*, which assumes that the color and intensity at a point's projection remain completely unchanged when the camera's viewpoint or the scene itself move in space. Unfortunately, while existing work on motion analysis and shape-from-stereo has relied almost exclusively on this constraint (Horn, 1986), it is valid only for static diffuse scenes. Our invariant extends recent attempts to generalize this constraint (Bérériat et al., 2000; Black et al., 2000; Haussecker and Fleet, 2000; Negahdaripour, 1998) so that shape recovery for non-diffuse scenes can also be studied in a rigorous manner.

In order to establish constraints on the temporal appearance variation of a scene point, we assume that the reflectance properties of every scene point remain unchanged as it moves through space. Unlike brightness constancy, which is violated even for diffuse scenes undergoing rotational motion, our *reflectance constancy assumption* only requires that the scene's physical properties remain unchanged during its motion. Given a moving scene point $\mathbf{p} = \hat{\mathbf{x}}(u_0, v_0, t)$ that is visible to the i-th camera and is illuminated by at least one light source at time t, this allows us to define a picture invariant $I_{i,t}^{\mathrm{Ainv}}(\mathbf{p})$ that is completely independent of viewpoint and time and can be thought of as an estimate of $\mathbf{p}$'s albedo:

Observation 2 (Dynamic Picture Invariant).

$$I_{i,t}^{\mathrm{Ainv}}(\mathbf{p}) \stackrel{\mathrm{def}}{=} \frac{I_i(\mathbf{p}) - I_i^{\mathrm{S}}(\mathbf{p})}{R_t^{\mathrm{D}}(\mathbf{p})} = \rho(u_0, v_0) = \text{constant}, \tag{8}$$

where

$$R_t^{\mathrm{D}}(\mathbf{p}) \stackrel{\mathrm{def}}{=} \sum_{l=1}^{L} C^{\mathrm{D}}(\mathbf{n}, \mathbf{l}_l - \mathbf{p}) \mathcal{L}_l(\mathbf{p}). \tag{9}$$

The Dynamic Picture Invariant of a point $\mathbf{p}$ is a simple "re-weighting" of the point's static invariant; its time-invariance follows directly from Observation 1 and Eqs. (2) and (5). Intuitively, by re-weighting the Static Picture Invariant we eliminate the invariant's dependence on surface orientation. The resulting quantity does not depend on viewpoint, surface orientation, or time under our reflectance constancy assumption.

3. The Dynamic Surfel Representation

Any general approach that attempts to recover 3D shape and motion by inverting the image formation model of Section 2 must explicitly take into account four properties of scene points—visibility, orientation, reflectance and motion. This is because these properties contribute simultaneously to the color and intensity of every pixel in the input views. We therefore

develop a global, spatially-distributed scene representation, called the *dynamic surfel representation*, that is specifically designed to make such an analysis possible. From a geometrical point of view, the dynamic surfel representation of a scene at time t is a sample-based description of the envelope of the scene's surface(s) at t (do Carmo, 1976), and the way that this envelope changes when the scene moves or deforms. It consists of a finite collection of *dynamic surfels* (surface elements), each of which describes the scene's tangent plane, motion, and picture invariants in a small neighborhood of a surface point:

Definition 1 (Dynamic Surfel). A dynamic surfel, $\mathcal{D}$, is represented by the tuple $\mathcal{D} \stackrel{\text{def}}{=} \langle t_0, \mathcal{S}, \mathcal{R}, \mathcal{B}, \mathcal{M} \rangle$, where t_0 is a time instant; $\mathcal{S}$ is the surfel's 3D shape component; $\mathcal{R}$ is its reflectance component; $\mathcal{B}$ is a parametric bump map; and $\mathcal{M}$ is its 3D motion component.

3.1. 3D Shape Component

The 3D shape component of a surfel is simply defined as the intersection of a plane and an ϵ-ball, $B(\mathbf{o}, \epsilon)$, that determines the surfel's spatial extent (Fig. 3(a)). We can therefore fully specify the surfel's shape component with seven parameters—three for the coordinates of the surfel's reference point $\mathbf{o}$, one for the size parameter ϵ, two for the surfel's unit normal $\mathbf{n}$, and one for the signed distance d of the surfel's plane from the reference point:

$$\mathcal{S} \stackrel{\text{def}}{=} \langle \mathbf{o}, \epsilon, \mathbf{n}, d \rangle. \tag{10}$$

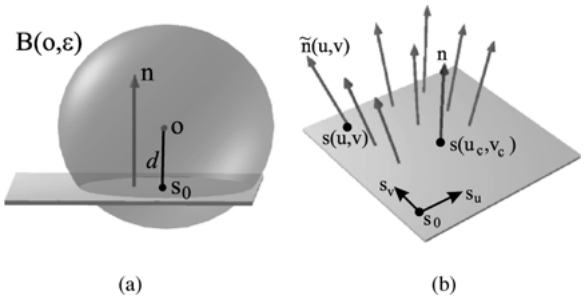

Figure 3. (a) The surfel 3D shape representation. The surfel's spatial domain is the circular intersection of the surfel plane with the ball $B(\mathbf{o}, \epsilon)$. When surfels are used to approximate non-planar scenes, this ball establishes a bound on the spatial extent of a surfel's planar approximation. (b) The surfel bump map representation, where $\mathbf{n}$ is the surfel plane's normal.

The surfel shape component induces a parameterization of all points on a surfel at a fixed instant $t = t_0$. This parameterization assigns a pair of coordinates to every surfel point. More specifically, let $\mathbf{s}_0$ be the surfel point closest to $\mathbf{o}$ and let $\mathbf{s}_u$, $\mathbf{s}_v$ be two arbitrary orthonormal vectors on the surfel's plane. The 3D coordinates of every point on a surfel are then uniquely determined by a pair of *surfel coordinates*, (u, v):

$$\mathbf{s}(u, v) \stackrel{\text{def}}{=} \mathbf{s}_0 + u\mathbf{s}_u + v\mathbf{s}_v. \tag{11}$$

Our surfel shape representation allows us to represent the scene's instantaneous global shape as a finite collection of surfels with shape components $\mathcal{S}_1, \ldots, \mathcal{S}_M$ (Fig. 12). This collection completely determines the visibility of all surfel points: if $\mathbf{s}$ is a point on an arbitrary surfel $\mathcal{S}_j$ in this collection and $\mathbf{c}_i$ is an input camera viewpoint or a light source position, we call $\mathbf{s}$ *visible from* $\mathbf{c}_i$ if and only if the open line segment $\mathbf{sc}_i$ does not intersect any surfel in the collection. Our surfel representation can therefore be used to model scenes that contain significant amounts of occlusion or self-shadows. In the following, we represent the visibility of a surfel point $\mathbf{s}$ with a binary function $\text{vis}(\mathbf{c}_i, \mathbf{s})$ that is equal to one if and only if $\mathbf{s}$ is visible from viewpoint $\mathbf{c}_i$.

3.2. Reflectance Component

We define the reflectance component of a surfel to be a two-parameter tuple that assigns a constant specular-lobe model β^{S} to every surfel point. More specifically, the reflectance model of Eq. (1) tells us that reflectance at each surfel point $\mathbf{s}(u, v)$ is completely determined by an albedo function $\rho(u, v)$ and the two functions $f(u, v)$ and $k(u, v)$ that specify the surfel's specular reflectance properties. We further constrain this model by requiring that the functions $f(u, v)$ and $k(u, v)$ be constant for all points on a surfel, and by allowing the constants f, k to vary arbitrarily from one surfel to the next:

$$\mathcal{R} \stackrel{\text{def}}{=} \langle f, k \rangle. \tag{12}$$

Our representation is therefore designed to model the reflectance of scenes whose specular component is either piecewise constant or varies slowly over the surface.

A key feature of this representation is that no albedo information is encoded in the representation itself.

From a mathematical point of view, this is because only the specular reflectance component is needed to compute Static and Dynamic Picture Invariants from the input views. Since our approach relies only on these invariants to relate pixels across multiple views and time instants, it allows us to recover a surfel-based scene representation without explicitly reconstructing the scene's albedo function, which may be very complex since it is allowed to vary arbitrarily over the scene's surface.

3.3. Parametric Bump Map

Bump maps are a popular computer graphics tool for improving the visual fidelity of synthesized images without increasing the geometric complexity of the underlying 3D scene representation (Blinn, 1978). They are typically used to convey an impression of surface roughness by altering the surface orientation attributes of individual points on a geometric model without altering the points' position in space. Here we associate a parametric bump map with every surfel to ensure that even though surfels are planar objects, they can still model appearance variations due to interactions between specular reflectance and high surface curvature.

Given a parameterization $\mathbf{s}(u, v)$ for the points on a surfel, the bump map replaces the point's original normal $\mathbf{n}$ in Eq. (6) with a modified normal, $\tilde{\mathbf{n}}(u, v)$ (Fig. 3(b)):

$$\tilde{\mathbf{n}}(u, v) \stackrel{\text{def}}{=} \mathbf{n} + [\mathbf{s}_u \quad \mathbf{s}_v] \begin{bmatrix} \kappa_{uu} & \kappa_{uv} \\ \kappa_{uv} & \kappa_{vv} \end{bmatrix} \begin{bmatrix} u - u_c \\ v - v_c \end{bmatrix}, \quad (13)$$

where $u_c, v_c, \kappa_{uu}, \kappa_{uv}, \kappa_{vv}$ are the parameters that define the bump map. This map effectively allows a quadratic variation in the surface normal across the surfel. This normal variation is consistent with that of a quadratic surface patch that touches the surfel's plane at a point with surfel coordinates (u_c, v_c). The three parameters $\kappa_{uu}, \kappa_{uv}, \kappa_{vv}$ control the apparent curvature of that patch and are known as the elements of the Gauss map differential, $d\mathbf{n}$ (do Carmo, 1976). Given a surfel parameterization $\mathbf{s}$, the bump map is determined by the five parameters specifying Eq. (13). This leads to a six-parameter description which explicitly encodes the surfel's parameterization in addition to the map itself:[6,7]

$$\mathcal{B} \stackrel{\text{def}}{=} \langle \mathbf{s}_u,\ u_c,\ v_c,\ \kappa_{uu},\ \kappa_{uv},\ \kappa_{vv} \rangle. \quad (14)$$

3.4. 3D Motion Component

To represent dynamic scenes, we augment our surfel shape representation with a smooth motion field that assigns an instantaneous 3D velocity to every surfel point. Recall that $\mathbf{s}(u, v)$ is the surfel's parameterization at time $t = t_0$. We describe the evolution of the surfel's shape in a small temporal neighborhood $[t_0 - \delta t, t_0 + \delta t]$ around t_0 by extending the surfel's static parameterization into the time domain:

$$\hat{\mathbf{s}}(u, v, t) \stackrel{\text{def}}{=} \mathbf{s}(u, v) + (t - t_0)\,(\hat{\mathbf{s}}_t + u\hat{\mathbf{s}}_{ut} + v\hat{\mathbf{s}}_{vt}), \quad (15)$$

where the time derivatives $\hat{\mathbf{s}}_t$, $\hat{\mathbf{s}}_{ut}$, and $\hat{\mathbf{s}}_{vt}$ are evaluated at time t_0. Intuitively, the vector $\hat{\mathbf{s}}_t$ captures the surfel's instantaneous translation while the vectors $\hat{\mathbf{s}}_{ut}$ and $\hat{\mathbf{s}}_{vt}$ capture all other motion-induced linear transformations of the surfel's plane. As such, the vectors $\hat{\mathbf{s}}_t, \hat{\mathbf{s}}_{ut}, \hat{\mathbf{s}}_{vt}$ can represent arbitrary translations, rotations, shearing and scaling of a surfel, but do not capture second-order effects due to changes in surface curvature. The resulting motion representation consists of the ten parameters defining these vectors and the neighborhood of t_0:

$$\mathcal{M} \stackrel{\text{def}}{=} \langle \delta t,\ \hat{\mathbf{s}}_t,\ \hat{\mathbf{s}}_{ut},\ \hat{\mathbf{s}}_{vt} \rangle. \quad (16)$$

Together, the shape, reflectance, bump map, and motion components of a surfel give rise to a 25-parameter planar and spatially-bounded scene representation with $|\mathcal{S}| = 7$, $|\mathcal{R}| = 2$, $|\mathcal{B}| = 6$ and $|\mathcal{M}| = 10$. As such, our surfel representation and, in particular, the bump map representation can be thought of as a compromize between the need to account for geometries, appearance variations and motions exhibited by non-planar scenes and the desire to minimize the number of required parameters and keep computations such as re-projection and image warping as simple as possible, i.e., modeled by homographies (Section 4).

4. Surfel Photo-Consistency

Our goal is to recover a surfel-based scene description from a set of input video sequences. In order to formalize this operation mathematically, we (1) rely on picture invariants to define the constraints that every valid solution to the reconstruction problem must satisfy, and (2) show that there always exists a surfel-based scene description that satisfies these constraints arbitrarily well.

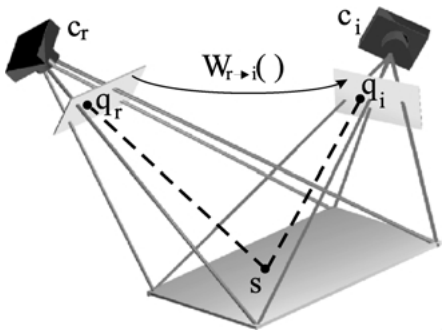

Figure 4. Pixel correspondences induced by a surfel point **s**. The warp function $\mathbf{W}_{r\to i}(\,)$ that maps pixels in the reference view to pixels along viewpoint $\mathbf{c}_i$ is a homography.

4.1. Static Photo-Consistency

Every point **s** on a 3D reconstruction of the scene induces a set of pixel correspondences between the input views in which it is visible (Fig. 4). Since the colors at **s**'s projection in the input views define a Static Picture Invariant when **s** is a point on the scene's true surface, we require that every point on a valid scene reconstruction satisfies this constraint as well.

In particular, let **s** be a surfel point that has an assigned normal $\tilde{\mathbf{n}}$, an assigned specular radiance $\beta^{\mathrm{S}}(\mathbf{s}, \tilde{\mathbf{n}}, \cdot, \cdot)$, and whose visibility from all cameras and light sources is known. Observation 1 tells us that in order to conform to the invariance constraints satisfied by the scene's true points, the color difference $I_i^{\mathrm{Dinv}} \stackrel{\mathrm{def}}{=} I_i(\mathbf{s}) - I_i^{\mathrm{S}}(\mathbf{s})$ must be constant for all viewpoints $\mathbf{c}_i$ where **s** is visible. This leads to the following definition for the photo-consistency of a surfel. Intuitively, this definition assesses the constancy of the Static Picture Invariant with respect to a reference view $\mathbf{c}_r$ in which the surfel is visible:

Definition 2 (Static Surfel Photo-Consistency). A surfel whose static properties are defined by the tuple $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$ is *photo-consistent* with the input views if and only if the error metric $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ is zero:

$$E_1(\mathcal{S}, \mathcal{R}, \mathcal{B}) \stackrel{\mathrm{def}}{=} \frac{1}{A_1} \int_S \sum_{i=1}^{N} \mathrm{vis}(\mathbf{c}_r, \mathbf{s})\, \mathrm{vis}(\mathbf{c}_i, \mathbf{s}) \times \left[I_i^{\mathrm{Dinv}}(\mathbf{s}) - I_r^{\mathrm{Dinv}}(\mathbf{s}) \right]^2 d\mathbf{s}, \quad (17)$$

where S is the set of 3D surfel points and A_1 is a normalizing factor equal to $\int_S \sum_{i=1}^{N} \mathrm{vis}(\mathbf{c}_r, \mathbf{s})\, \mathrm{vis}(\mathbf{c}_i, \mathbf{s})\, d\mathbf{s}$.

Since there is a one-to-one correspondence between surfel points and pixels in the input views, the invariant metric $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ can be expressed in terms of the pixels at the projection of a surfel. This re-formulation allows us to evaluate $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ without having to explicitly sample 3D points on the surfel itself:

$$E_1(\mathcal{S}, \mathcal{R}, \mathcal{B}) = \frac{1}{A_1} \int_Q \sum_{i=1}^{N} \mathrm{vis}(\mathbf{c}_r, \mathbf{q})\, \mathrm{vis}(\mathbf{c}_i, \mathbf{W}_{r\to i}(\mathbf{q})) \times \left[I_i^{\mathrm{Dinv}}(\mathbf{W}_{r\to i}(\mathbf{q})) - I_r^{\mathrm{Dinv}}(\mathbf{q}) \right]^2 d\mathbf{q}, \quad (18)$$

where Q is the set of pixels in the surfel's footprint along view $\mathbf{c}_r$; $\mathbf{W}_{r\to i}(\,)$ is the warp function that maps pixels in $\mathbf{c}_r$ to pixels in $\mathbf{c}_i$ (Fig. 4); $\mathrm{vis}(\cdot, \mathbf{q})$ is defined in terms of the visibility of the 3D point projecting to pixel **q**; and A_1 is equal to $\int_Q \sum_{i=1}^{N} \mathrm{vis}(\mathbf{c}_r, \mathbf{q})\, \mathrm{vis}(\mathbf{c}_i, \mathbf{W}_{r\to i}(\mathbf{q}))\, d\mathbf{q}$.

Since the shape of a surfel is always planar, the pixel correspondences it induces between the input views can be expressed as homographies (Appendix A). The metric $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ can therefore be thought of as the sum-of-square-differences of pixels in a reference picture-invariant image I_r^{Dinv}, and in its homography-warped counterparts, I_i^{Dinv}, $i = 1, \ldots, N$. Moreover, since these patches are identical to the input images when the scene is known to be Lambertian (i.e., when $\beta^{\mathrm{S}} = 0$), the metric is essentially a generalization of warp-based metrics that have been previously proposed for 2D image registration (Irani and Peleg, 1991; Szeliski, 1996; Kanatani and Ohta, 1999), motion estimation (Szeliski, 1999; Irani et al., 1997), and stereo (Faugeras and Keriven, 1998; Loop and Zhang, 1999) in Lambertian settings.

4.2. Dynamic Photo-Consistency

The notion of dynamic photo-consistency for 3D points and surfels follows as a straightforward generalization of the static photo-consistency constraint. In particular, let **q** be the projection of a 3D point **s** at time t_0 along a viewpoint $\mathbf{c}_r$, and let $\mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q})$ be the pixel corresponding to **q** at time t from viewpoint $\mathbf{c}_i$. Observation 2 tells us that in order to conform to the scene's picture-invariant properties, every pixel $\mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q})$ must define a color $I_{i,t}^{\mathrm{Ainv}}$ that is constant across all viewpoints and all time instants where **s** is visible. This leads to the following metric for quantifying the dynamic photo-consistency of an entire surfel:

Definition 3 (Dynamic Surfel Photo-Consistency). A surfel $\mathcal{D} = \langle t_0, \mathcal{S}, \mathcal{R}, \mathcal{B}, \mathcal{M} \rangle$ is *photo-consistent* with the input views if and only if the error metric $E_2(\mathcal{S}, \mathcal{R}, \mathcal{B}, \mathcal{M})$ is zero:

$$E_2(\mathcal{S}, \mathcal{R}, \mathcal{B}, \mathcal{M}) \stackrel{\text{def}}{=} \frac{1}{A_2} \int_T \int_Q \sum_{i=1}^{N} \text{vis}(\mathbf{c}_r, \mathbf{q}) \, \text{vis}\big(\mathbf{c}_i, \mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q})\big) \times \big[I_{i,t}^{\text{Ainv}}\big(\mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q})\big) - I_{r,t_0}^{\text{Ainv}}(\mathbf{q})\big]^2 d\mathbf{q}\, dt, \quad (19)$$

where Q is the set of pixels on the surfel's footprint along view $\mathbf{c}_r$ at time t_0; the interval $T = [t_0 - \delta t, t_0 + \delta t]$ is defined by the surfel's temporal extent, δt; $\mathbf{W}_{r\to i}^{t_0\to t}(\ )$ is the warp function that maps pixels in $\mathbf{c}_r$ at time t_0 to pixels in $\mathbf{c}_i$ at time t; and A_2 is equal to $\int_T \int_Q \sum_{i=1}^{N} \text{vis}(\mathbf{c}_r, \mathbf{q}) \, \text{vis}(\mathbf{c}_i, \mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q}))\, d\mathbf{q}\, dt$.

A useful consequence of our dynamic surfel representation is that the space of 3D motions representable by a surfel's motion component is identical to the space of plane-to-plane homographies. This allows us to use a common computational framework for assessing static and dynamic photo-consistency, based on appropriately-defined homography warps.

4.3. *Sufficiency of the Surfel-Based Scene Representation*

While the photo-consistency metrics defined above tell us when a surfel is photo-consistent, they do not tell us anything about whether photo-consistent surfels can indeed be found for every scene or scene point. The answer to this question is not obvious because a surfel-based scene description is only an approximation of the scene's true shape, reflectance, and motion properties. Hence, even though the metrics E_1 and E_2 are identically zero when evaluated at points on the scene's true surface(s), they are not zero, in general, when evaluated at non-surface points. Importantly, a negative answer to the question would imply that our surfel-based representation scheme is not powerful enough to generate photo-consistent reconstructions of arbitrary scenes. Fortunately, the following theorem establishes the sufficiency of our scheme in the static case. Intuitively, Theorem 1 says that we can almost always define a surfel whose deviation from photo-consistency falls within an arbitrarily-small error bound δ:

Theorem 1 (*Surfel Approximation Theorem*). *For every scene point* $\mathbf{p}$ *that does not project to a shadow boundary or an occlusion boundary there exists a surfel shape component* $\mathcal{S} = \langle \mathbf{o}, \epsilon, \mathbf{n}, d \rangle$ *and components* $\mathcal{R}, \mathcal{B}$ *such that* $\mathbf{p}$ *is contained in the ball* $B(\mathbf{o}, \epsilon)$ *and* $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B}) < \delta$.

See Appendix B for a proof as well as a theorem that gives a precise description of what the spatial extent ϵ must be to achieve consistency for a given scene (Theorem 4). The Surfel Approximation Theorem is important in our analysis for three reasons. First, it tells us that individual surfels can reproduce to an arbitrary degree of consistency the scene's appearance in the neighborhood of almost every scene point.[8] Second, because the theorem's proof is constructive, it provides a detailed study of the inter-relationship between the photo-consistency error bound δ, the scene's local surface geometry at $\mathbf{p}$, the scene's albedo function, and the surfel's spatial extent ϵ. As such, it makes explicit the class of surfels that can represent a specific scene neighborhood in a way that preserves photo-consistency, and the class of scenes that can be reconstructed from a given family of surfels. Third, it suggests that we can construct a global photo-consistent representation of the scene by defining a discrete collection of surfels, each of which approximates the scene's surface(s) in a well-defined spatial neighborhood.

While in theory we can make the spatial extent of a surfel arbitrarily small to ensure photo-consistency, in practice a surfel's footprint in the input views must contain enough pixels to allow estimation of its 25 parameters. We consider the problem of building such surfel-based scene approximations in the next section and present experimental results with real scenes in Section 8.

5. Shape and Reflectance by Surfel Sampling

At the heart of our approach lies the problem of computing a set of surfels that cover the scene's visible surfaces. Since the scene's shape is unknown, we must answer three questions: (1) how do we identify the regions of space that contain surface points, (2) how do we determine the cameras and light sources that reach those regions, and (3) how do we use the input images to fit surfels to these regions?

Let $\mathcal{V}^{\text{init}}$ be a known and finite volume that contains the scene to be reconstructed, and let $B_1, \ldots, B_V$ be a finite set of ϵ-balls whose union is $\mathcal{V}^{\text{init}}$. To answer the

above questions, we reduce global shape computation to the problem of performing a *space query* in a ball $B = B(\mathbf{o}, \epsilon)$ and introduce a computational framework called *surfel sampling* to perform the query:

Definition 4 (Space Query). (1) Determine whether B intersects the scene's volume and (2) if it does, compute a normal $\mathbf{n}$, a distance d, and reflectance and bump map components $\mathcal{R}, \mathcal{B}$ so that the surfel defined by $\mathcal{S} = \langle \mathbf{o}, \epsilon, \mathbf{n}, d \rangle$ and $\mathcal{R}, \mathcal{B}$ is photo-consistent with the input views.

Space queries require assessing a surfel's photo-consistency, an operation which requires knowing the visibility relationship between points in B and the cameras and light sources. In the following we assume that the visibility function $\text{vis}(\cdot, \mathbf{s})$ is known for every point $\mathbf{s} \in B$ and focus on how to answer the first and third questions we posed above. We return to the question of how these visibilities can be determined in Section 7.

5.1. *Space Queries by Surfel Sampling*

The set of all surfels in a ball $B(\mathbf{o}, \epsilon)$ is an 11-dimensional subset of the space of shape, reflectance, and bump map components. Surface sampling conducts an organized exploration of this space in search of the *globally optimal surfel*, $\langle \mathcal{S}^*, \mathcal{R}^*, \mathcal{B}^* \rangle$, i.e., the surfel in B that minimizes the static photo-consistency metric, E_1:

$$\langle \mathcal{S}^*, \mathcal{R}^*, \mathcal{B}^* \rangle \stackrel{\text{def}}{=} \arg\min_{\substack{\mathbf{n} \\ d<\epsilon}} \Big[\min_{\mathcal{R},\mathcal{B}} E_1(\mathcal{S}, \mathcal{R}, \mathcal{B}) \Big], \quad (20)$$

with $\mathcal{S} = \langle \mathbf{o}, \epsilon, \mathbf{n}, d \rangle$. Once this surfel is identified, the value of the photo-consistency metric is used to either reject the ball as empty space (i.e., B does not intersect the scene volume), or to accept the optimal surfel as a valid local approximation of the scene points in B. The Surfel Approximation Theorem of Section 4.3 tells us that such a global minimization inside $B(\mathbf{o}, \epsilon)$ is sound mathematically, i.e., it is guaranteed to generate to a photo-consistent scene approximation if the ball intersects the scene and ϵ is sufficiently small.

Querying space by minimizing E_1 is difficult for three reasons. First, a local minimization of E_1 is not sufficient to decide if B is empty because, in practice, the value of E_1 at a local minimum is *not* a good predictor of its value at the global minimum (Fig. 5). Intuitively, this is because even a small deviation from

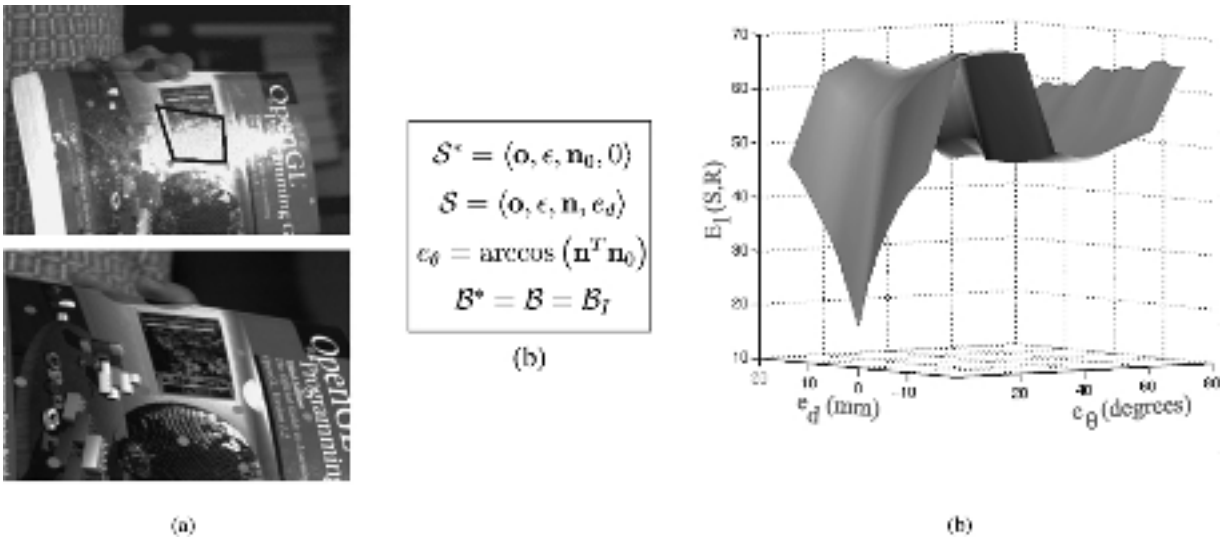

Figure 5. Behavior of the metric $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ in the neighborhood of a planar scene patch, outlined in black, on the book shown in (a). Note that the book cover exhibits a strong specularity for one of the two input cameras shown, due to the configuration of the scene's two light sources. (b) The true 3D position and orientation of the patch was measured in advance to obtain a ground-truth value for $\mathcal{S}^*$. The bump map $\mathcal{B}_I$ represents the identity map, i.e., the map that assigns the surfel plane's normal, $\mathbf{n}$, to every point on the surfel. (c) A plot of E_1 with $\mathcal{R}$ computed according to Section 5.2, with 100 sample points per surfel ($P = 100$). Note that the global minimum of E_1 occurs at $\mathcal{S} = \mathcal{S}^*$ and is sharply lower than the rest of the function, suggesting that global minimization of E_1 will lead to accurate shape recovery despite the patch's strongly specular appearance. Also note the deep local valley far from $\mathcal{S}^*$, which suggests that global minimization of E_1 is difficult even when all but two of the surfel's shape parameters are known exactly.

a surfel's optimal position and orientation will corrupt stereo correspondences and specularity computations, making it hard to compute a picture invariant from the input images. Second, the metric exhibits deep local minima in practice, making it impossible to guarantee global minimization using standard methods (e.g., Levenberg-Marquardt (Press et al., 1988)). Third, an exhaustive search of $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-space is practically impossible because of its high dimensionality.

To overcome these difficulties, the Surfel Sampling Algorithm combines a coarse sampling of $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-space along specific dimensions with a sequence of linear and non-linear optimization steps that "explore" the neighborhood of each $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-sample. The algorithm's strategy is to rely as much as possible on linear optimization to generate and screen an initial set of "seed" samples since it is computationally much less expensive than non-linear optimization or explicit sampling.

A graphical depiction of the Surfel Sampling Algorithm is shown in Fig. 6. The order and mathematical formulation of the minimization steps are of fundamental importance to the method because they determine the "size" of the neighborhood that can be explored from a single sample. By choosing them appropriately we can therefore minimize the number of dimensions that have to be explicitly sampled as well as the density of the samples themselves. We consider each of these steps below.

5.2. *Linear Reflectance Estimation*

Steps 1–4 of the Surfel Sampling Algorithm are directed toward efficiently searching for $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-samples that may be near the globally optimal surfel in $B(\mathbf{o}, \epsilon)$. This is done by (1) generating a sample $\mathcal{S}$ of linear shape parameters, (2) generating a value for the specular exponent, k, (3) using a linear method to compute an assignment for the surfel's specular coefficient, f, and (4) rejecting all resulting $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-samples that produce high values for the photo-consistency metric E_1.

More specifically, we observe that the scene image formation model described by Eq. (4) becomes linear when the surfel's shape component, specular exponent, and bump map are known. For a point $\mathbf{s} = \mathbf{s}(u_0, v_0)$

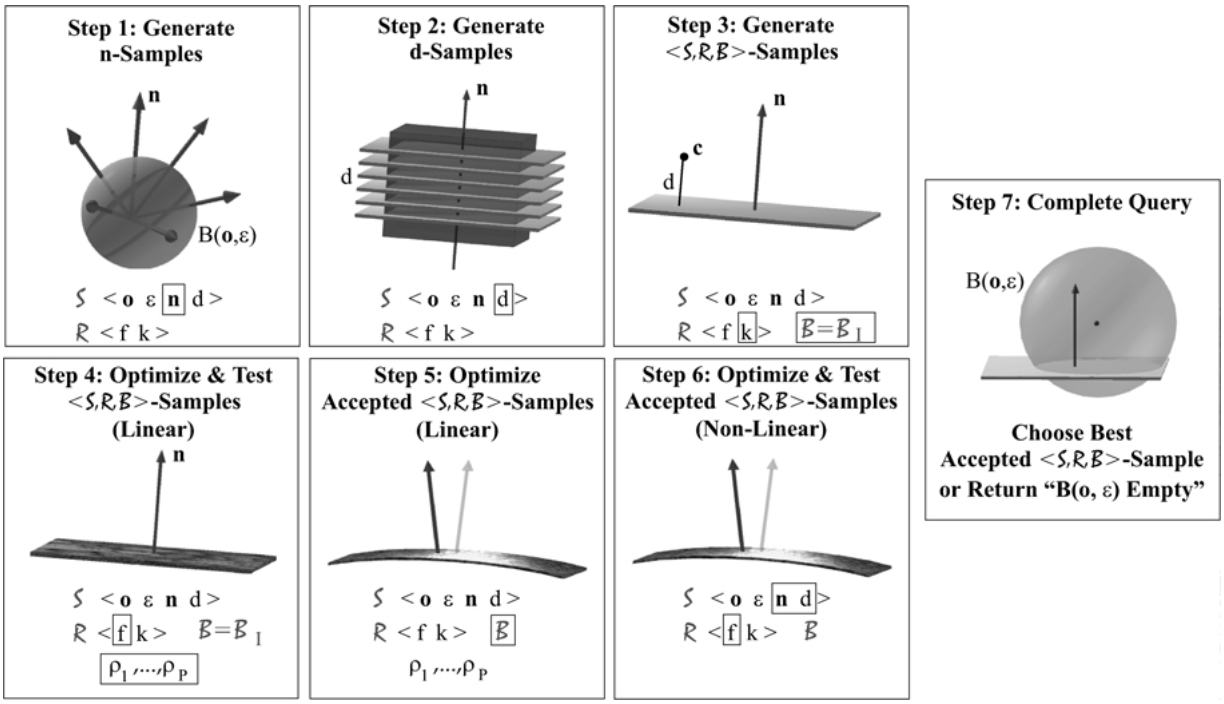

Figure 6. The Surfel Sampling Algorithm. Boxed parameters indicate the parameters to which sampling or optimization is applied. The optimization Steps 4, 5, and 6 are described in Sections 5.2, 5.3 and 5.4, respectively. Note that Step 4 computes a set of discrete albedo values for the surfel as a by-product of the step's linear optimization method. These values are used for optimizing the surfel's bump map in Step 5 but are not part of the surfel's representation and are not used for measuring surfel photo-consistency.

on the surfel, this observation leads to a linear equation with only two unknowns—the point's albedo, $\rho(u_0, v_0)$, and the surfel's specular coefficient, f:

$$I_i(\mathbf{s}) = a\rho(u_0, v_0) + bf, \tag{21}$$

where $I_i(\mathbf{s})$ is the pixel color at $\mathbf{s}$'s projection in the i-th camera and a, b collect the known terms of Eq. (4). For P surfel points projecting to a total of Q pixels in the input views, Eq. (21) gives rise to a linear system of Q equations and $P+1$ unknowns, corresponding to f and the albedo values, $\rho_1, \ldots, \rho_P$, of the individual surfel points. In practice, we form the system by assigning an identity bump map to the surfel and uniformly sampling the surfel's parameterization, $\mathbf{s}(u, v)$, in the interior of the ball B. This results in a sparse linear system that is solvable in $O(Q)$ steps.

Since $\mathcal{S}$ and k must be known to formulate the linear system, we generate them through sampling. We first uniformly sample the 2D space of surfel normals so that neighboring normals form an angle that is less than a constant ψ, and then uniformly sample the 1D space of distances, d, from the center of $B(\mathbf{o}, \epsilon)$. These two sampling steps span the entire space of planes in B (Steps 1 and 2 in Fig. 6). A coarse sampling of the space of specular exponent values completes the set of parameters needed for linear reflectance estimation. In our experiments, reported in Section 8, a total of about 20,000 $\langle \mathcal{S}, \mathcal{R} \rangle$-samples are generated for each space query. Each one of these samples is then accepted for further refinement only if its photo-consistency is among the K-best in the generated sample set, where K is a pre-determined constant (usually between 5 and 10).

5.3. Bump Map Estimation

Linear reflectance estimation assumes that scene points in B can be approximated by a plane. Ignoring curvature when the scene in B is curved will lead to an incorrect (i.e., sub-optimal) surfel solution since photo-consistency cannot be ensured. This is especially important for non-diffuse scenes, where the strong effect of surface curvature on the appearance of specularities is well known (Blake and Bulthoff, 1991). To overcome this difficulty we note that when the orientation of a surfel and the positions of the light sources are known, we can reason directly about the presence or absence of specularities and about how, by interacting with surface curvature, they affect a surfel's appearance.

We use this idea in two ways. First, we slightly modify the linear reflectance estimation step of Section 5.2 to make it robust to curvature-induced effects. This is accomplished by estimating the surfel albedos $\rho_1, \ldots, \rho_P$ from a subset of its input views, i.e., those views where curvature-induced effects due to a strong specular highlight cannot be present because of the relation between their viewpoint, the surfel normal, and the light source positions (Fig. 7(a)). Second, we assign a bump map to every surfel that exhibits a specular highlight in at least one input view. This allows us to build photo-consistent surfel-based reconstructions of curved, specular scenes that would not have been otherwise possible[9] (Section 8).

Our bump map estimation procedure is based on the following theorem. Let $\mathcal{S} = \langle \mathbf{o}, \epsilon, \mathbf{n}, d \rangle$ be the shape component of a surfel and let $\rho_1, \ldots, \rho_P$ be the known albedos of P surfel points $\mathbf{s}_1, \ldots, \mathbf{s}_P$, respectively. Theorem 2 establishes an explicit relation between the color at the projection of a point $\mathbf{s}_j$ and the surfel's optimal bump map parameters, i.e., the parameters that reproduce this color exactly:

Theorem 2 (*Linear Bump Map Estimation Theorem*)**.** *Suppose that* (1) *the scene's specular exponent is constant in the ball* $B(\mathbf{o}, \epsilon)$ *and has a known value* k, (2) *the contribution of specular reflectance to every color* $I_i(\mathbf{s}_j)$ *in the* i*-th view is negligible for all but one light source* $\mathbf{l}_l$, *and* (3) *the bump map origin, defined by surfel coordinates* (u_c, v_c), *is chosen so that the ray through* $\mathbf{s}(u_c, v_c)$ *and* $\mathbf{c}_i$ *is along the direction of perfect specular reflection for light source* $\mathbf{l}_l$. *The difference between color* $I_i(\mathbf{s}_j)$ *and the color predicted by the surfel's bump map and reflectance models is zero for all* $j = 1, \ldots, P$ *if and only if the following equation is satisfied for all such* j

$$\left(\mathbf{v}_c^{\text{out}}\right)^T \begin{bmatrix} \kappa_{uu} & \kappa_{uv} \\ \kappa_{uv} & \kappa_{vv} \end{bmatrix} \mathbf{A}_{jc} \begin{bmatrix} \kappa_{uu} & \kappa_{uv} \\ \kappa_{uv} & \kappa_{vv} \end{bmatrix} \left(\mathbf{v}_c^{\text{in}}\right) - b_{ij}\, f^{-\frac{1}{k}} = -1 + e(\epsilon), \tag{22}$$

where

$$\mathbf{A}_{jc} = 2 \begin{bmatrix} u_j - u_c \\ v_j - v_c \end{bmatrix} [u_j - u_c \quad v_j - v_c],$$

$$\mathbf{v}_c^{\text{out}} = \begin{bmatrix} \mathbf{s}_u{}^T \\ \mathbf{s}_v{}^T \end{bmatrix} \mathbf{d}_c^{\text{out}}, \quad \mathbf{v}_c^{\text{in}} = \begin{bmatrix} \mathbf{s}_u{}^T \\ \mathbf{s}_v{}^T \end{bmatrix} \mathbf{d}_c^{\text{in}},$$

$$b_{ij} = \left[\frac{I_i(\mathbf{s}_j) - I_i^{\text{Dinv}}(\mathbf{s}_j)}{\mathcal{L}_l(\mathbf{s}_j)} \right]^{\frac{1}{k}},$$

$$\lim_{\epsilon \to 0} e(\epsilon) = 0,$$

$$\mathbf{s}_j = \mathbf{s}(u_j, v_j),$$

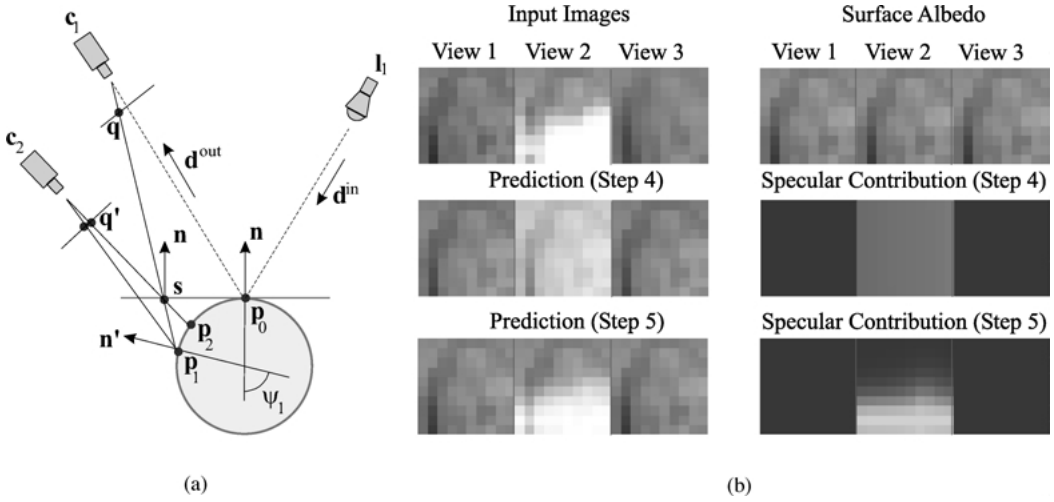

Figure 7. Curvature compensation. (a) Errors due to planar approximations of a curved scene. The static picture invariant at **s**'s projections will be erroneously estimated for two reasons. First, these projections are projections of two distinct scene points $\mathbf{p}_1, \mathbf{p}_2$. Second, the normal of scene point $\mathbf{p}_1$ is $\mathbf{n}'$, not $\mathbf{n}$. For the specular viewpoint $\mathbf{c}_1$, the prediction error in Eq. (4) is dominated by the error in $\mathbf{n}$. If $\mathbf{p}_1$ has a large specular exponent, however, the actual contribution of specular reflectance to $\mathbf{q}$'s color will be negligible. To increase the robustness of estimating the albedo samples, $\rho_1, \ldots, \rho_P$, we ignore near-specular views of a surfel, i.e., those where $\arccos[C^S(\mathbf{n}, \mathbf{d}^{\text{out}}, \mathbf{d}^{\text{in}})] < \psi_{\min}$ for at least one surfel point and one light source. The angle $\psi_{\min}$ is determined from the surfel's (known) specular exponent. (b) Appearance prediction results with and without curvature compensation for the specular "vase" scene in Fig. 25. Shown are three views of a 10×10 set of samples from a single surfel (left column) and the predicted contributions to these views from each of the surfel's reflectance components (right column). Note the significant specular color errors in Step 4, where the surface normal is the same for all surfel points.

and $\mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}}$ *are the unit orientations of the rays* $\mathbf{s}(u_c, v_c)\mathbf{c}_i$ *and* $\mathbf{s}(u_c, v_c)\mathbf{l}_l$, *respectively.*

See Appendix C for a proof. Intuitively, Theorem 2 tells us that when the scene generates a strong specular highlight at a viewpoint $\mathbf{c}_i$ and the surfel we use to approximate the scene is sufficiently small, we can compute a near-optimal bump map by solving a linear system with unknowns κ_{uu}^2, $\kappa_{uu}\kappa_{uv}$, κ_{uv}^2, $\kappa_{uv}\kappa_{vv}$, κ_{vv}^2, $\kappa_{vv}\kappa_{uu}$ and $f^{-\frac{1}{k}}$. The only requirement is that the surfel's position, orientation and specular exponent approximate those of the true scene. In practice, we obtain this system with the help of the previously-computed surfel shape and reflectance parameters $\mathcal{S}, \mathcal{R}$ and the albedo values, $\rho_1, \ldots, \rho_P$.

5.4. Non-Linear Shape and Reflectance Estimation

While our linear estimation methods assign reflectance and bump map components to a surfel of known shape $\mathcal{S}$, the surfel's position and orientation is determined strictly through sampling. Hence, the density of samples in $\langle\mathcal{S}\rangle$-space determines the degree to which a surfel can approximate the shape of the true scene. This leads to sub-optimal surfel solutions. In order to further refine the parameters of a computed $\langle\mathcal{S}, \mathcal{R}, \mathcal{B}\rangle$-sample without densely sampling $\langle\mathcal{S}\rangle$-space, we use the sample $\langle\mathcal{S}, \mathcal{R}, \mathcal{B}\rangle$ as a starting point in a non-linear minimization stage. This stage relies on Levenberg-Marquardt's optimization algorithm to minimize the static photo-consistency metric E_1 over all surfel parameters except those defining the specular exponent k and the bump map $\mathcal{B}$.[10]

The $\langle\mathcal{S}, \mathcal{R}, \mathcal{B}\rangle$-sample returned from the non-linear minimization stage is accepted as a candidate solution to the space query if $E_1(\mathcal{S}, \mathcal{R}, \mathcal{B})$ is less than a bound δ^{nlin}. This bound can be determined by taking into account image noise, the sampling densities in Steps 1–3 of the Surfel Sampling Algorithm, and the scene reconstructibility conditions that result from Theorem 1 (Appendix B). Once all such candidates have been identified, the candidate with the smallest photo-consistency error is returned as the solution to the space query. If no candidates exist, i.e., the metric E_1 is larger than δ^{nlin} for all computed surfels, the ϵ-ball defined by the query is considered to be empty of scene points.

6. Dynamic Surfel Reconstruction

Any general approach to the problem of reconstructing dynamic scenes must inevitably account for the

very complex interactions between 3D shape, appearance, motion and illumination (e.g., changing shape, moving shadows and specularities, dynamic illumination effects due to changes in surface orientation, moving occlusion boundaries, etc). As a first step toward this goal, we use an approach that is based on two basic principles. First, even though the above interactions are complex and non-linear, they *can* be resolved when an estimate is available for the scene's instantaneous 3D shape and reflectance. Second, temporal variations in a scene's appearance do not only constrain the 3D motion of the scene—they strongly constrain the scene's instantaneous 3D shape as well. We apply these two principles for dynamic surfel recovery by (1) using a recovered surfel-based description of the scene's instantaneous global shape to identify all cameras, light sources, and input pixels that can contribute to a surfel's 3D motion estimate, (2) developing a new direct linear method that uses this information to assign a motion component, $\mathcal{M}$, to an $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$-sample, and (3) using an additional, non-linear estimation step that jointly refines $\mathcal{S}$, $\mathcal{R}$ and $\mathcal{M}$ to ensure the dynamic photo-consistency of the resulting dynamic surfel.

6.1. Direct Linear 3D Motion Estimation

Our linear estimation approach is based on the observation that when a point on a Lambertian scene is not on a shadow boundary or an occlusion boundary and when the point is distant from the scene's light sources, only one factor can affect the point's dynamic appearance—a change in its surface orientation. Here we generalize this observation in order to handle scene points that are not diffuse and in order to recover their translational and non-translational 3D motion components. To achieve this, we relate a point's 3D motion to the variation of its Static Picture Invariant in a given view, which is equal to the point's color in the Lambertian case.

Specifically, suppose $\mathbf{p}(t) = \hat{\mathbf{x}}(u_0, v_0, t)$ is a moving surface point that does not project to a shadow boundary or an occlusion boundary at time t_0. The total time derivative of $\mathbf{p}$'s Static Picture Invariant, $I_i^{\text{Dinv}}(\mathbf{p})$, in view i satisfies the following two equations:

$$\frac{d}{dt} I_i^{\text{Dinv}}(\mathbf{p}) = \rho(u_0, v_0)\, \mathbf{d}(\mathbf{p})^T \frac{\partial \mathbf{n}}{\partial t} + \rho(u_0, v_0) \frac{\partial \mathbf{d}}{\partial t}(\mathbf{p})^T \mathbf{n} \tag{23}$$

$$\frac{d}{dt} I_i^{\text{Dinv}}(\mathbf{p}) = \left[\frac{\partial}{\partial \mathbf{p}} I_i^{\text{Dinv}}(\mathbf{p}) \right] [\hat{\mathbf{x}}_t + u_0 \hat{\mathbf{x}}_{ut} + v_0 \hat{\mathbf{x}}_{vt}] + \frac{\partial}{\partial t} I_i^{\text{Dinv}}(\mathbf{p}), \tag{24}$$

where the total derivative is evaluted at t_0, partials are evaluated at (u_0, v_0, t_0), and $\mathbf{d}(\mathbf{p})$ is given by

$$\mathbf{d}(\mathbf{p}) = \sum_{l=1}^{L} \mathcal{L}_l(\mathbf{p}) \frac{\mathbf{l}_l - \mathbf{p}}{\|\mathbf{l}_l - \mathbf{p}\|}. \tag{25}$$

Equation (23) is obtained by substituting Eq. (2) into Eq. (5) and differentiating the result with respect to time. It tells us that if we subtract the contribution of specular reflections from $\mathbf{p}$'s projection in the i-th camera, the resulting color change can have only two causes—a change in $\mathbf{p}$'s surface orientation or a change in its position relative to the light sources. Equation (24) goes a step further, allowing us to relate the color and intensity variations at a specific image pixel to the 3D translation of $\mathbf{p}$ and to the deformation and reorientation of the scene in $\mathbf{p}$'s neighborhood. It generalizes the optical flow constraint equation (Horn, 1986) in order to capture the effects of changes in surface orientation relative to the light source(s) and to account for the scene's non-Lambertian reflectance.

To use Eqs. (23) and (24), we concentrate on the case where the inter-frame translation of point $\mathbf{p}$ is much smaller than its distance from the light sources:[11]

Observation 3 (Motion-Induced Variation of the Static Picture Invariant). If all of the scene's light sources are located at infinity and defined by the unit vectors $\mathbf{d}_1^{\text{in}}, \ldots, \mathbf{d}_L^{\text{in}}$, the following equality holds:

$$\rho(u_0, v_0)\, \mathbf{d}^{\infty}(\mathbf{p})^T \frac{\partial \mathbf{n}}{\partial t} = \left[\frac{\partial}{\partial \mathbf{p}} I_i^{\text{Dinv}}(\mathbf{p}) \right] [\hat{\mathbf{x}}_t + u_0 \hat{\mathbf{x}}_{ut} + v_0 \hat{\mathbf{x}}_{vt}] + \frac{\partial}{\partial t} I_i^{\text{Dinv}}(\mathbf{p}) \tag{26}$$

with

$$\mathbf{d}^{\infty}(\mathbf{p}) = \sum_{l=1}^{L} \mathcal{L}_l(\mathbf{p}) \mathbf{d}_l^{\text{in}}.$$

Observation 3 follows from Eqs. (23) and (24) by noting that $\mathbf{d}^{\infty}(\mathbf{p})$ is the limit of $\mathbf{d}(\mathbf{p})$ when all light sources are at infinity and noting that the temporal derivative of $\mathbf{d}^{\infty}(\mathbf{p})$ in Eq. (23) is zero. The observation leads directly to the following theorem which allows us to formulate the recovery of a surfel's 3D motion parameters as a direct linear estimation problem. In particular, let $\mathbf{p}_j$, $j = 1, \ldots, P$ be scene points with

known albedos ρ_j that are contained on a planar surface region with normal $\mathbf{n} = \mathbf{x}_u \wedge \mathbf{x}_v$ at time t_0, and that have coordinates (u_j, v_j) with respect to the surface parameterization $\hat{\mathbf{x}}$:

Theorem 3 (*Linear 3D Motion Estimation Theorem*)**.** *The vectors $\hat{\mathbf{x}}_t, \hat{\mathbf{x}}_{ut}, \hat{\mathbf{x}}_{vt}$ describing the plane's 3D motion satisfy the $Q \times 9$ system*

$$\begin{bmatrix} \vdots \\ \mathbf{a}_{ij}^T \\ \vdots \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_t \\ \hat{\mathbf{x}}_{ut} \\ \hat{\mathbf{x}}_{vt} \end{bmatrix} = \begin{bmatrix} \vdots \\ -\frac{\partial}{\partial t} I_i^{\mathrm{Dinv}}(\mathbf{p}_j) \\ \vdots \end{bmatrix} + \mathbf{e}(t), \qquad \lim_{t \to t_0} \mathbf{e}(t) = \mathbf{0}_{Q \times 1}, \quad (27)$$

where Q is the total number of pixels to which the points $\mathbf{p}_j$ project, and $\mathbf{a}_{ij}$ is the row contributed to the system from the projection of $\mathbf{p}_j$ in the i-th camera:

$$\mathbf{a}_{ij} \stackrel{\mathrm{def}}{=} \left[\frac{\partial}{\partial \mathbf{p}_j} I_i^{\mathrm{Dinv}}(\mathbf{p}_j) \quad u_j \frac{\partial}{\partial \mathbf{p}_j} I_i^{\mathrm{Dinv}}(\mathbf{p}_j) + \rho_j \left\{ [\mathbf{d}^{\infty}(\mathbf{p}_j) \wedge \mathbf{x}_v]^T + \mathbf{d}^{\infty}(\mathbf{p}_j)^T \mathbf{n} \mathbf{x}_u^T \right\} \quad v_j \frac{\partial}{\partial \mathbf{p}_j} I_i^{\mathrm{Dinv}}(\mathbf{p}_j) + \rho_j \left\{ [\mathbf{x}_u \wedge \mathbf{d}^{\infty}(\mathbf{p}_j)]^T + \mathbf{d}^{\infty}(\mathbf{p}_j)^T \mathbf{n} \mathbf{x}_v^T \right\} \right]. \quad (28)$$

See Appendix D for a proof. Theorem 3 shows that we can recover a surfel's 3D motion parameters directly from the Static Picture Invariants in the input views by treating the surfel as a local approximation of the scene's true 3D surface, defining P points on the surfel's plane and computing their albedos, and solving the linear system in Eq. (27) with $\mathbf{e}(t) = \mathbf{0}_{Q \times 1}$.

Besides allowing us to compute 3D motion estimates, Theorem 3 has two important practical consequences. First, it allows us to compute a surfel's motion by simultaneously integrating all pixels in all input views to which the surfel projects. In general, when a surfel is visible from multiple cameras the system will be highly over-determined, making the estimation less sensitive to image noise. Second, it allows us to use singular value decomposition (SVD) to identify those 3D motion parameters of a surfel that can be estimated reliably from the input views. This is particularly useful since real scenes frequently contain regions where motion information cannot be extracted because of the aperture problem (Horn, 1986).

In practice, we use Theorem 3 to compute a surfel's motion with the help of a coarse-to-fine estimation algorithm that consists of six steps: (1) sample P points on a surfel and compute their albedos (Sections 5.2 to 5.4), (2) compute the Static Picture Invariants defined by consecutive frames of the input videos (Section 2.3), (3) build the Gaussian pyramids (Burt and Adelson, 1983) of these images, (4) use SVD analysis on the linear system of Eq. (27) to evaluate which, if any, of the surfel's motion parameters can be reliably estimated,[12] (5) solve the system for those parameters at level h of the pyramid, and (6) refine the motion parameters by repeating these steps for level $h - 1$.

The above algorithm can be thought of as both a generalization and a restriction of previous work on physically-based 2D flow estimation (Haussecker and Fleet, 2000; Negahdaripour, 1998). On one hand, existing physically-based approaches for recovering the motion of specular and/or diffuse scenes attempt to recover a parametric 2D flow field from a single video stream, a problem that is inherently under-constrained. This leads to flow fields that are consistent with the input sequence but do *not* describe the scene's true 3D motion. Unlike these techniques, our multi-view generalization leads to a well-posed 3D estimation problem whose solution, as indicated by Theorem 3, corresponds to the scene's true 3D motion. On the other hand, our linear estimation approach works by effectively discarding the contribution of specular reflectance from the input views before 3D motion estimation is performed. Since prominent specularities are strong cues about the scene's 3D shape and motion, this information ought to be used in the shape and motion estimation process, as in existing 2D techniques (Negahdaripour, 1998), rather than discarded. This is a topic of future work.

6.2. *Non-Linear 3D Motion, Shape and Reflectance Estimation*

An increasing body of evidence suggests that joint estimation of 3D shape and motion parameters can significantly improve shape and motion computations (Vedula et al., 2000; Carceroni and Kutulakos, 1999a, 1999b). This is because temporal variations due to 3D motion provide additional shape constraints that cannot be captured by stereo and reflectance information alone. Here we apply this principle by jointly optimizing a surfel's 3D shape, reflectance and motion. To achieve this, we apply Levenberg-Marquardt optimization to minimize the dynamic photo-consistency

metric, $E_2(\mathcal{S}, \mathcal{R}, \mathcal{B}, \mathcal{M})$ (Section 4.2). The minimization is performed over the parameters of the surfel's 3D shape component, $\mathcal{S}$, the surfel's specular factor, f, and those parameters of its 3D motion component that were determined to be reliable through SVD analysis.

7. Global Scene Reconstruction

Our space query formalism and its spatio-temporal extension have two main characteristics. First, they are designed to solve a local reconstruction problem, i.e., that of computing the shape, reflectance and motion properties of the scene in a small 4D space-time neighborhood. Second, they assume that the visibility of all points inside that neighborhood is known with respect to the input cameras and light sources. To generate a global 4D scene description we therefore need to answer three questions: (1) how can we use space queries to compute a global instantaneous description of the scene's shape and reflectance at every time instant, (2) how can we resolve the instantaneous visibility of 3D points, and (3) how can we recover a global instantaneous 3D motion field that describes the scene's motion at every time instant? We answer these questions within the context of a volumetric reconstruction framework based on space carving (Kutulakos and Seitz, 2000).

Let $\mathcal{V}^{\text{init}}$ be a known and finite volume that contains the scene as an unknown sub-volume. We represent $\mathcal{V}^{\text{init}}$ as a finite collection of voxels $v_1, \ldots, v_V$ and use space queries to determine whether or not a voxel in this collection intersects the scene volume. For computational simplicity, we restrict the spatial extent of a space query to be a single voxel v_m in this collection and use space queries to fit at most one surfel to v_m. Using this representation, instantaneous shape reconstruction consists of repeatedly applying space queries to individual voxels in the volume, and leads to reconstructions that contain at most V distinct surfels.

In order to compute the visibility function $\text{vis}(\cdot, \mathbf{p})$ for every point $\mathbf{p}$ inside a voxel, we observe that we can compute an approximation to this function by applying space queries to $\mathcal{V}^{\text{init}}$'s voxels in a specific order. This observation, initially exploited in Kutulakos and Seitz (2000), Seitz and Dyer (1999), Szeliski and Golland (1998) and Langer and Zucker (1994), is based on the idea that if we iteratively "carve away," voxels from $\mathcal{V}^{\text{init}}$, we guarantee that the visibility set of all surface points on the remaining volume is monotonic, i.e., that the set of cameras (or light sources) from which a point is visible can only increase. Here we use this observation by (1) carving from $\mathcal{V}^{\text{init}}$ every voxel whose space query determines that it does not intersect the scene, (2) applying a space query only to voxels on the surface of the uncarved subset in $\mathcal{V}^{\text{init}}$, (3) approximating $\text{vis}(\mathbf{c}_i, \mathbf{p})$ for a point $\mathbf{p}$ in a voxel by a binary function that is one if and only if the segment $\mathbf{c}_i\mathbf{p}$ does not intersect an uncarved voxel or a previously-computed surfel,[13] and (4) updating this function after each carving operation.[14]

To reconstruct the scene's instantaneous 3D motion at every time t, we use the instantaneous description of the scene's shape and reflectance to assign a motion component to every reconstructed surfel according to Section 6. The above considerations lead to the following algorithm for recovering a surfel-based description of a dynamic scene from multiple views:

Dynamic Surfel Reconstruction Algorithm

Step 1. Initialize $\mathcal{V}^{\text{init}}$ to a volume containing the true scene.

Step 2. (Instantaneous Shape & Reflectance Recovery for every frame t)

Step 2a. Initialize the volume, $\mathcal{V} = \mathcal{V}^{\text{init}}$, and the surfel collection, $\Sigma_t = \{\ \}$.

Step 2b. Repeat the following steps for voxels $v_m \in$ *Surface*($\mathcal{V}$) until no voxels are carved away:

- For every 3D point $\mathbf{p}$ in v_m, define $\text{vis}(\mathbf{c}_i, \mathbf{p})$ to be one if and only if the line segment $\mathbf{c}_i\mathbf{p}$ does not intersect a voxel in $\mathcal{V} - \{v_m\}$ or a surfel in Σ_t.
- If there are 3D points in v_m that are visible by at least two cameras:
  1. Set $\mathcal{V} = \mathcal{V} - \{v_m\}$.
  2. Assign to v_m the reference camera r that maximizes v_m's visible projected area (i.e., the area of v_m's projection not occluded by any voxels in $\mathcal{V}$ or surfels in Σ_t).
  3. Perform a space query on v_m; if the query returns a sample $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle$, set $\Sigma_t = \Sigma_t \cup \{\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle\}$.

Step 3. (Instantaneous Shape, Reflectance & Motion Recovery for every frame t). Repeat the following steps for every tuple $\langle \mathcal{S}, \mathcal{R}, \mathcal{B} \rangle \in \Sigma_t$:

Step 3a. Compute $\mathcal{S}$'s visibility from the light sources at t and $t + 1$. If this visibility changes between the two frames, do not estimate the surfel's motion and continue with a new surfel.

Figure 8. Our 7-camera acquisition rig (cameras circled).

Step 3b. Compute the set of cameras from which $\mathcal{S}$ is visible in both frames t and $t+1$.

Step 3c. Compute the Static Picture Invariants $I_{i,t}^{\text{Dinv}}$ and $I_{i,t+1}^{\text{Dinv}}$ and the Dynamic Picture Invariants $I_{i,t}^{\text{Ainv}}$ and $I_{i,t+1}^{\text{Ainv}}$ in the neighborhood of $\mathcal{S}$'s projection.

Step 3d. Use the above information to compute a dynamic surfel $\langle \mathcal{S}', \mathcal{R}', \mathcal{B}, \mathcal{M} \rangle$ that assigns a motion component $\mathcal{M}$ to the surfel and refines the surfel's shape and reflectance components.

8. Experimental Results

To demonstrate the applicability of our approach we performed experiments with a number of complex, dynamic real scenes. Multi-view sequences were acquired with a rig of seven synchronized, progressive-scan Pulnix TCM-9700 color cameras (Fig. 8). The cameras allowed simultaneous observation of an approximately $30 \times 30 \times 30$ cm working volume. A sequence of geometric and radiometric calibration steps ensured that the projection of points within the working volume was accurate to approximately 0.5 pixels and that color and intensity agreement between cameras was on the order of 1–5 gray levels per channel. Scene illumination consisted of two point light sources approximately 3 m away from the scene, whose 3D positions were recovered by adapting the method in Bouguet and Perona (1998). We used color images for all computations, treating each band as an independent image.

Our method was applied to a variety of 7-view sequences ranging from 10 to over 100 frames. The four examples shown here were chosen to illustrate its performance for scenes of dramatically different shape, reflectance, and motion dynamics: (1) the "T-shirt" sequence involves the manual deformation of a densely-textured, near-Lambertian and fairly thick fabric (Fig. 9), resulting in a very smooth stretching

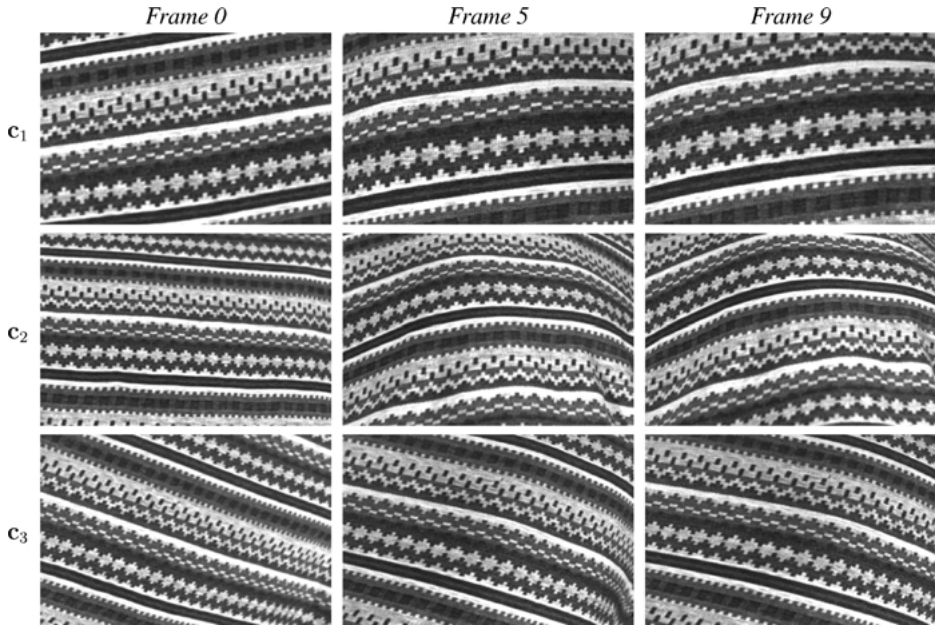

Figure 9. The "T-shirt" sequence: a T-shirt is "punched" from below, undergoing an upward motion and deformation. Three out of ten frames and three out of seven views are shown. Each row corresponds to a different camera viewpoint.

deformation of the shirt's surface and no self-shadows or self-occlusions; (2) the "flag" sequence involves the rapid shaking of a very thin near-Lambertian flag with both highly-textured and sparsely-textured regions (Fig. 11)—the flag's low internal cohesion induced a very dramatic 3D shape change that caused the emergence of self-occlusions and self-shadows, involved little or no stretching deformation, and induced a motion field that varied significantly over the flag's surface in both magnitude and direction; (3) the "neck" sequence illustrates the approach's behavior on images of the body, where the surface has non-Lambertian reflectance and contains many regions with little texture, and where self-occlusions and complex deformations occur during the neck's motion (Fig. 21); and (4) the "vase" sequence involves the rigid motion of a highly-specular object (Fig. 25).

In all of the above examples, the working volume was divided equally into a $16 \times 12 \times 8$ array of cells for performing space queries. Since every cell can have at most one surfel describing it, this limited the "resolution" of our reconstructions to a maximum of 1408 surfels. Only two parameters of the algorithm were changed across sequences—minor adjustments to the bounding box and adjustments to the variance threshold for surfel rejection. We relied on the Phong reflectance model for all sequences (Section 2.2) and used exactly the same rates for sampling surfel space. Specifically, orientation sampling (Step 1 of the Surfel Sampling Algorithm) was achieved by choosing 193 uniformly-spaced samples of the Gaussian hemisphere that faced the cameras. This corresponded to approximately 10 degrees between neighboring samples for the surfel normal. The distance between z-samples was 0.5 mm (Step 2 of the Surfel Sampling Algorithm), and the same four discrete values (1.5, 6, 25 and 100) were used for the cosine lobe exponent, k.

8.1. Shirt Sequence

Figure 10 shows reconstruction results for the "shirt" sequence. From the point of view of shape recovery, the scene's Lambertian reflectance and dense texture can be thought of as representing a best-case scenario for traditional stereo techniques. Our results suggest that the Surfel Sampling Algorithm performs well on this sequence too. Three observations can be made about these results. First, in addition to providing a set of raw 3D points as in most existing stereo techniques, our reconstructions provide explicit information about surface orientation which is not always easy to extract from a set of noisy 3D points (Amenta et al., 1998). Second, the recovered shapes display a great degree of smoothness and global consistency even though each surfel was recovered completely independently and no smoothness or regularization criteria were used in this process. Intuitively, the surfel representation induces this coherence by ensuring that each recovered surfel explains a fairly large set of pixels in the input views, ranging from 200 to 1000 pixels per surfel per image. This coherence occurs to a great degree in all of our reconstructions, suggesting that smoothness constraints and global shape models are not always needed for extracting globally-consistent shapes from images. Third, the recovered 3D motion fields suggest that a scene's 3D motions can be very complex—our spatially-decomposed surfel representation allows 3D reconstruction of motion fields whose local direction and magnitude can vary significantly over the surface while still being globally consistent.

8.2. Flag Sequence

Results from the "flag" sequence (Fig. 11) are shown in Figs. 12 and 13. Even though the scene is textured and near-Lambertian, it is challenging both for traditional stereo and for recent scene-space stereo techniques (e.g., space carving (Kutulakos and Seitz, 2000)). In particular, the self-occlusions occurring at many time instants during the sequence create multi-view image sets where the visibility of individual scene points changes dramatically from camera to camera. This makes it difficult to recover the scene's changing shape through time without explicitly reasoning about the occlusion relationships of the many input views. While recent scene-space stereo algorithms have been shown to handle such cases successfully, their reliance on matching pixels one by one through a simple color comparison test makes the "flag" sequence an almost worst-case scenario for these methods—the small number of input views and the many image regions with uniform or slowly-varying colors make color-based correspondence finding an inherently ambiguous process, leading to reconstructions that are overly conservative and highly non-smooth (Fig. 14). From the point of view of static shape recovery, our approach therefore incorporates into a scene-space stereo framework the spatial coherence constraints and region-based correlation metrics found in traditional binocular stereo techniques (Ohta and Kanade, 1985).

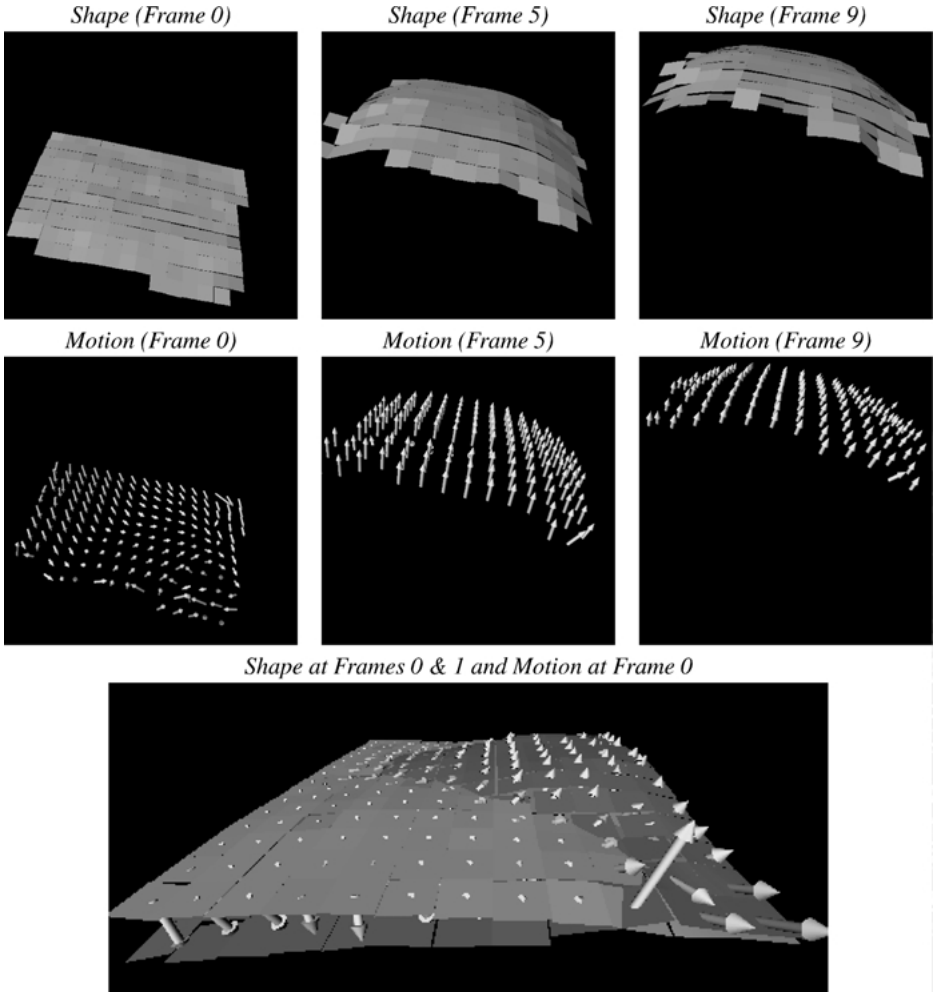

Figure 10. Reconstruction results for the "T-shirt" sequence.

While the static reconstruction results in this sequence illustrate our method's utility in recovering shape, a key feature of the method is its ability to capture the highly-complex 3D motion fields generated by surfaces such as clothing and skin. These motion fields are recovered *at sub-surfel resolution*, since each surfel is assigned a local parametric motion field that can vary over its surface (Fig. 12). Note that it would be difficult to recover such fields by instrumenting the flag with sensors or reflectors (such as those used for human 3D motion capture) since this instrumentation would likely affect the flag's physical properties and, hence, its motion dynamics.

To better evalute the accuracy of our reconstructions in the absence of ground truth, we relied on a technique that is commonly used to assess errors in image alignment and motion estimation (Caspi and Irani, 2000): we used the 3D shape and 3D motion information computed for frames t, $t+1$ to perform view transfer across viewpoints and through time, and compared these predictions to images from the input sequences. Figure 15 shows three input images, $I_{1,4}$, $I_{1,5}$, $I_{3,5}$, corresponding to camera $\mathbf{c}_1$ and Frame 4, camera $\mathbf{c}_1$ and Frame 5, and camera $\mathbf{c}_3$ and Frame 5, respectively. The instantaneous 3D shape computed at Frame 5 was used to warp image $I_{3,5}$ in order to predict the flag's appearance from a different input view at that instant. Additionally, the instantaneous 3D motion field computed at Frame 4 was used to "unwarp" images at Frame 5 in order to negate the scene's 3D motion and "stabilize" views of the scene. A similar set of results for Frames 14 and 15 is shown in Fig. 16.

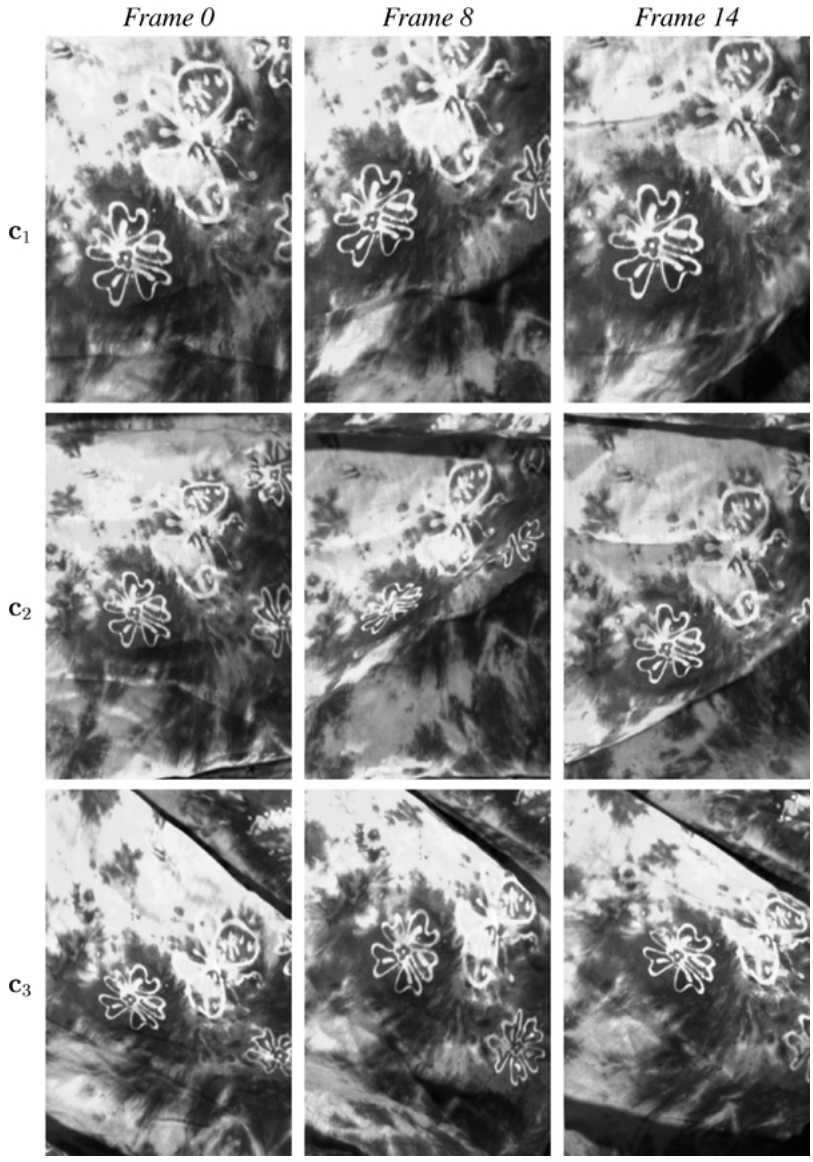

Figure 11. The "flag" sequence: a hanging rectangular cloth is shaken briskly from its top two corners. Three out of forty frames and three out of seven views are shown. Note that the cloth self-occludes from viewpoint $\mathbf{c}_2$, at Frame 14.

Figure 17 shows the difference images created by subtracting the predictions of Fig. 15 from the corresponding input views. For comparison purposes, the figure also shows difference images obtained by subtracting raw input views from different viewpoints and/or time instants, i.e., without performing view transfer or motion compensation. These results suggest that while some errors still exist, image variations due to camera position and motion are accounted for quite accurately by the computed 3D shape and 3D motion estimates. Similar results are shown in Fig. 18 for the predictions of Fig. 16. A plot of the computed prediction errors for several time instants is shown in Fig. 19.

Two observations can be made from our results on the flag sequence. First, our motion compensation results (e.g., first column of Figs. 17 and 18) suggest that the computed 3D motions induce a 2D flow field in each input view that is sufficiently accurate to "undo" the scene's apparent motion from that view. Second, even though overall prediction errors are quite low for both motion compensation and view transfer, the transfer

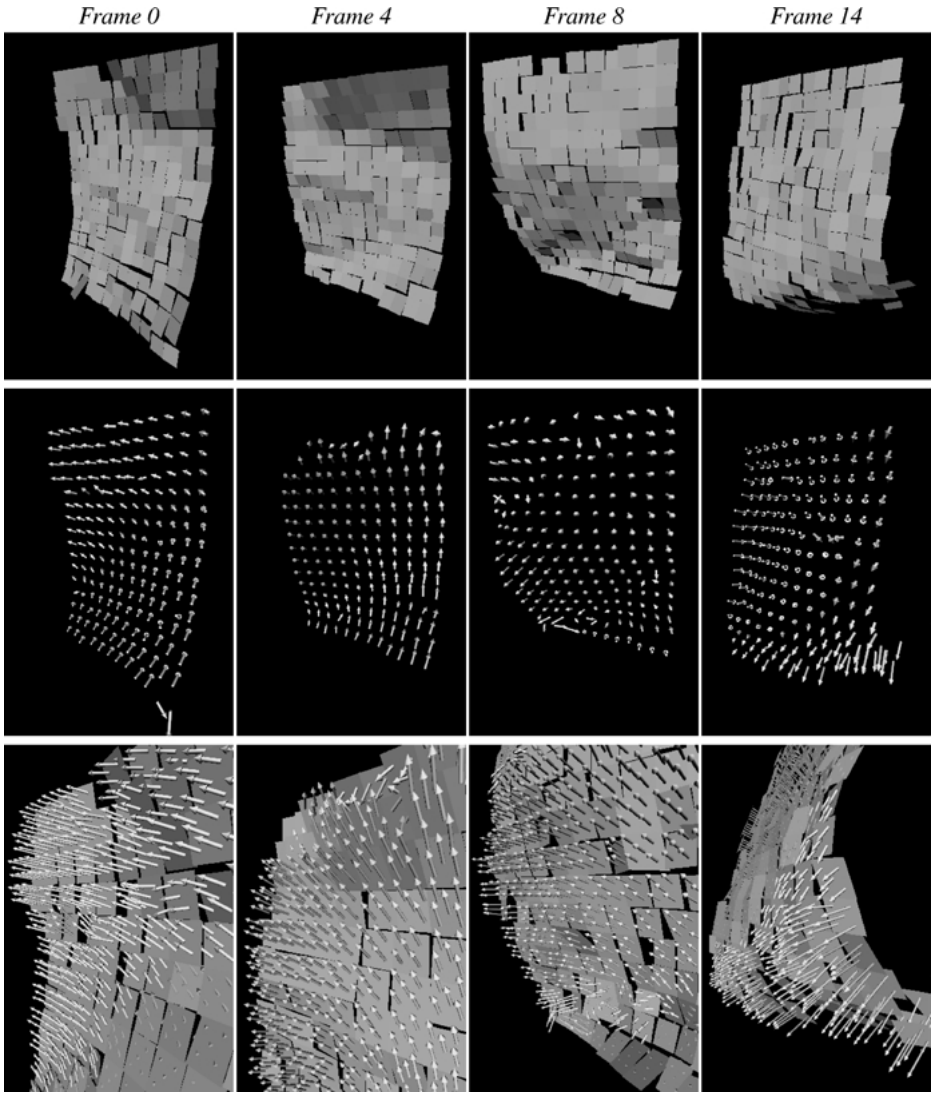

Figure 12. Reconstruction results for "flag" sequence. *Top row*: Reconstructed surfels. *Middle row*: Reconstructed 3D motion field. One vector is drawn per surfel. *Bottom row*: Close-up views of surfels and motion field with five 3D motion vectors drawn per surfel. Note the intra-surfel motion variations in the motion field reconstructed for Frame 14.

of images from one viewpoint to another (e.g., second column of Figs. 17 and 18) causes noticeably larger prediction errors than motion compensation for a single viewpoint (e.g., first column of Figs. 17 and 18). These errors appear as "edges" in the difference images, suggesting that some high frequency information is lost during the view transfer process. A closer examination of the input views suggests that a major source of these errors is the fact that the input cameras do not capture the scene's appearance with an equal level of detail (Fig. 20). This is because differences in the cameras' distance from the object as well as the cameras' field-of-view induce a view-dependent blurring of the scene's appearance, which is not captured by our image formation model.[15] Importantly, these results suggest that our approach is sufficiently robust to accurately recover the scene's shape despite this view-dependent blurring that cannot be accounted for in our model.

8.3. Neck Sequence

Results on recovering the motion field of a complex, deforming skin surface (Fig. 21) are shown in Figs. 22

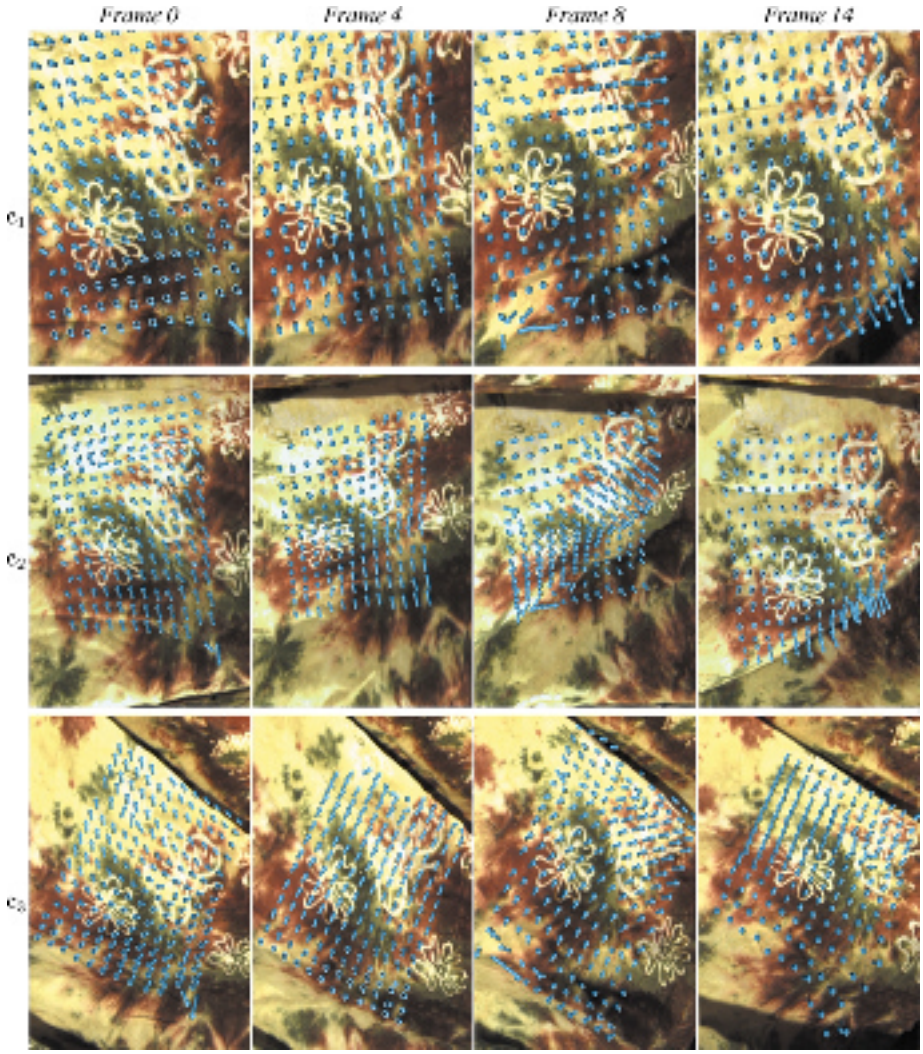

Figure 13. Reconstructed 3D motion field, viewed from the position of the input cameras and overlaid with the original "flag" sequence. The field is reconstructed only for the scene that intersects the user-defined working volume. Note how the motion field recovered for Frame 14 is not affected by the self-occlusions along viewpoint $\mathbf{c}_2$: as shown in the view along $\mathbf{c}_1$, the field is reconstructed in its entirety, even for the regions that are occluded from $\mathbf{c}_2$.

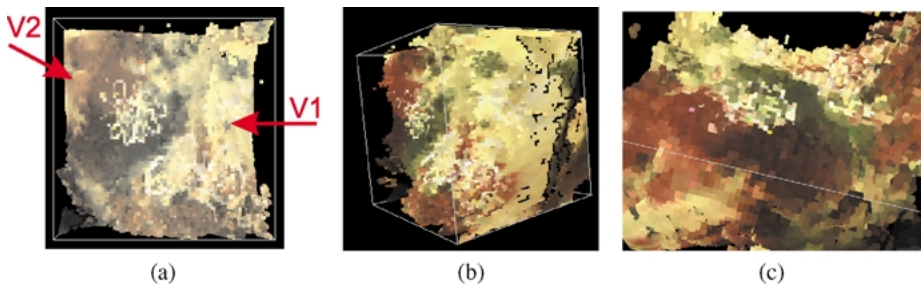

Figure 14. Reconstruction results using the Space Carving Algorithm (Kutulakos and Seitz, 2000). The algorithm was initialized to the same bounding box as the Surfel Sampling Algorithm. (a) Face-on view of the cloth. The white pattern in the middle of the view corresponds to the "flower" that appears prominently in Fig. 13. (b), (c) Views of the reconstruction along the arrows $V1$, $V2$ in (a), respectively. Note that the flower appears severely distorted in these views.

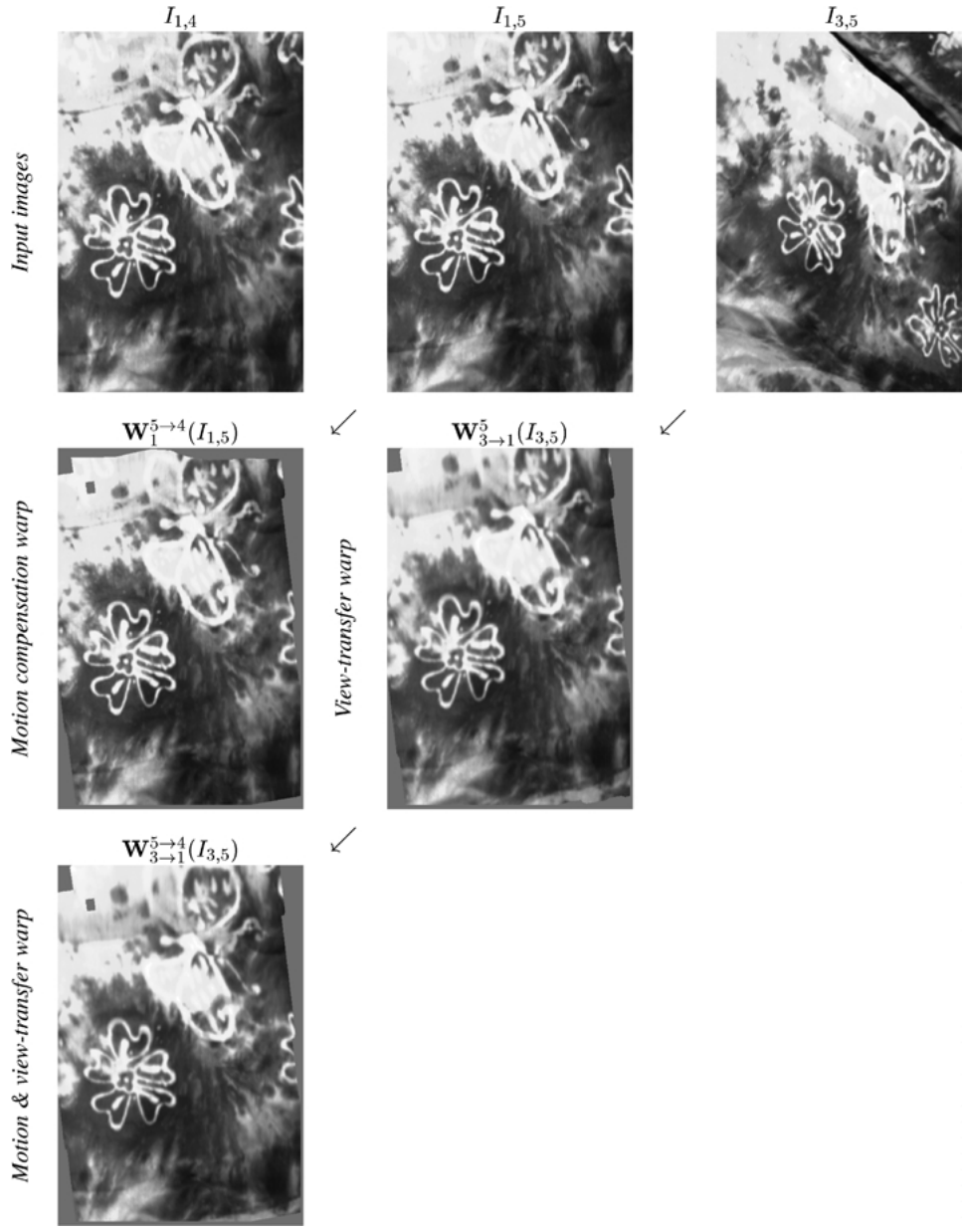

Figure 15. Results of using the 3D shape and motion estimates in Fig. 12, Column 2 for view transfer and motion compensation. *Top row*: Input images. *Center and bottom rows*: The computed 3D shape and motion estimates define global 2D warp fields, $\mathbf{W}_{i\to j}^{t\to t'}$, that allow transfer of an input image I from view $\mathbf{c}_i$ and frame t to a new view $\mathbf{c}_j$ and a new frame t'. The uniformly-colored areas near the edges of the images correspond to areas that are not covered by the footprint of any surfel. Note that images along a single column in the figure should be identical—with the exception of those areas mentioned above—if the 3D shape and motion estimates contain no errors.

and 23. Unlike the fairly small deformations occurring on the face due to facial expressions (DeCarlo and Metaxas, 1998), the neck surface moves significantly during head motion, creating new surface features and self-occlusions, and causing deformations that vary considerably from region to region (e.g., near the mouth vs. near the chest). We are not aware of existing techniques that can capture such 3D motions reliably.

Figure 24 shows 3D shape reconstruction results for three instants of the neck sequence. The results show that, with a few exceptions, the reconstructed surfel

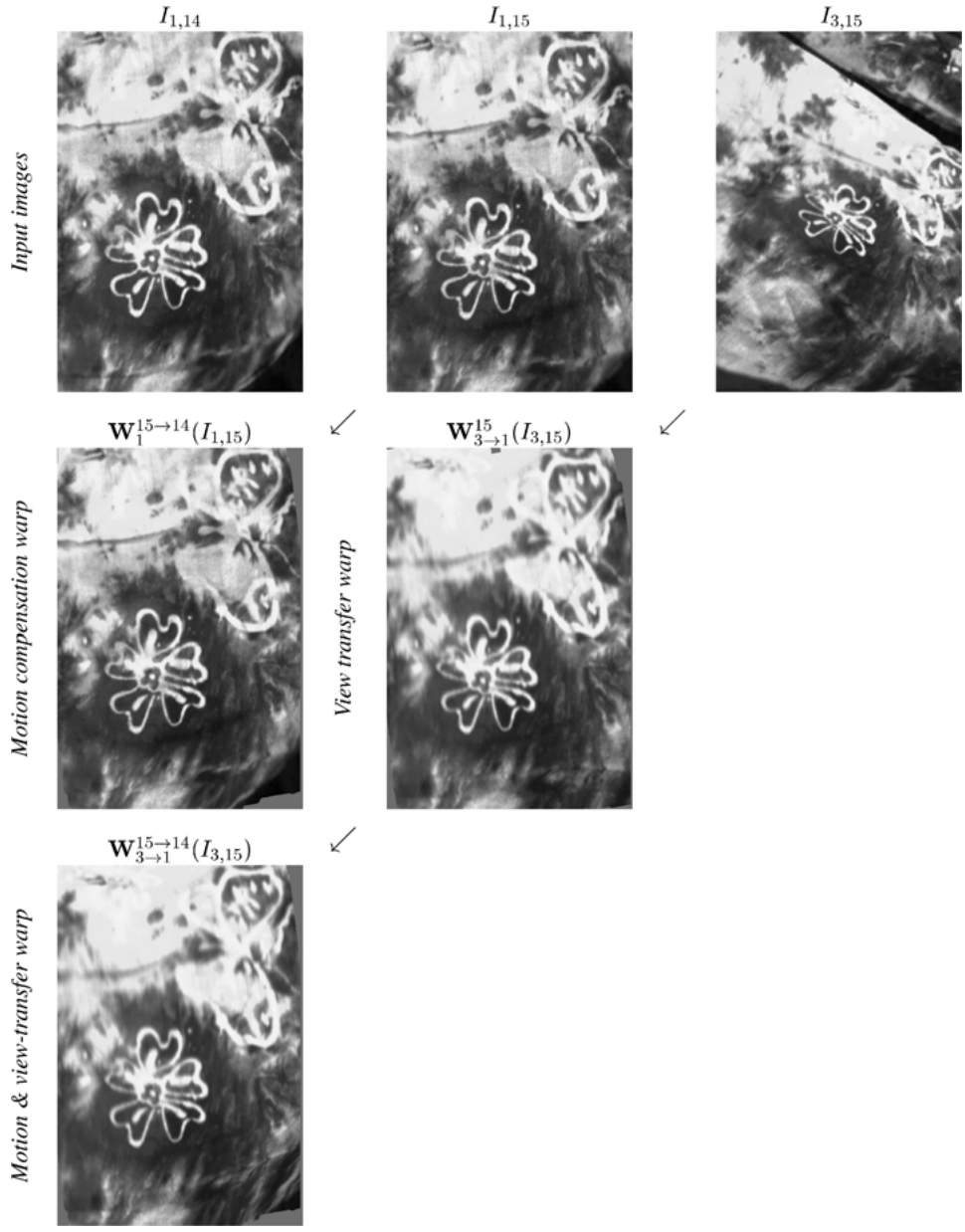

Figure 16. Results of using the 3D shape and motion estimates in Fig. 12, Column 4 for view transfer and motion compensation.

representation captures both the 3D position and the orientation of the neck surface quite well, including areas with fine surface structure and little texture information. Failures of the technique include (1) "stray" surfels that do not correspond to any scene surface, (2) regions that are recovered with incorrect orientation, and (3) regions that have not been approximated by a surfel. As in previous scene-space algorithms, the Surfel Sampling Algorithm is prone to incorrectly reconstruct surfels whose projected footprints are inside a uniformly-colored Lambertian region in all views. While such regions need to contain hundreds of pixels to affect a surfel's correct reconstruction, they do occur—the dark shirt and dark background contributed significantly to the creation of such surfels, as did some regions on the neck itself. Un-reconstructed regions occur because of gaps between individually-reconstructed surfels or because no photo-consistent

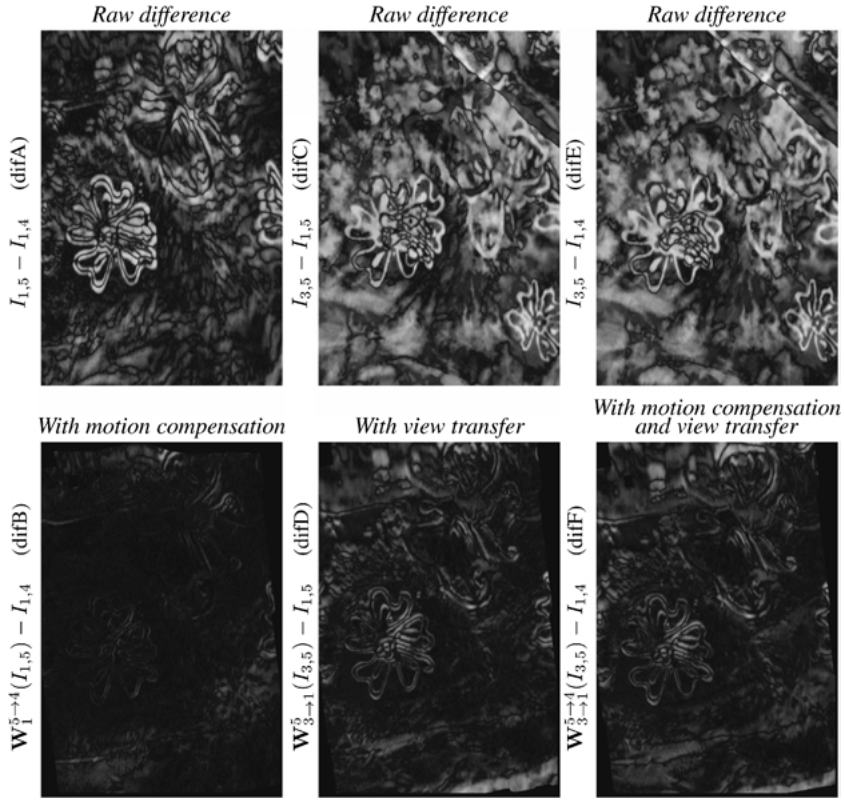

Figure 17. Intensities of pairwise differences between images in Fig. 15. The overall accuracy of our shape and motion computations is captured by the difference images in the right-most column.

surfel was found. Gaps between surfels can be filled by performing additional space queries that are in the vicinity of already-reconstructed surfels and are not aligned with the volumetric grid. Moreover, since our surfel representation can be thought of as a sparse reconstruction of the scene's envelope, it is possible to fill gaps by computing a set of 3D surfaces that are tangent to the computed surfels and are consistent with the input views. This topic is currently under investigation.

8.4. Vase Sequence

Shape, reflectance and motion reconstruction results for the "vase" scene (Fig. 25) are shown in Figs. 26, 7(b) and 27, respectively. Despite the scene's strong specular highlights in some of the input views, the scene's shape and its 3D motion field were recovered to a great degree. The key reason for this behavior is our method's ability to reason about the existence of highlights and account for them during motion processing. As Fig. 27 shows, failure to model such highlights explicitly leads to corrupted 3D motion estimates. This illustrates the importance of our bump map representation which, as shown in Figs. 27 and 7(b), allows us to model the curvature-induced variations in the appearance of a highlight that occur *within* a surfel's projection.

The results of Fig. 26 also demonstrate our method's ability to recover good shape estimates with surfels of different sizes. This suggests that our basic algorithm can be applied repeatedly in a coarse-to-fine manner in order to refine shape estimates and increase the resolution of the computed motion fields. The development of such an algorithm is beyond the scope of this paper and is a topic of current work.

9. Concluding Remarks

This paper introduced *surfel-based reconstruction* as a new, general mathematical framework for recovering

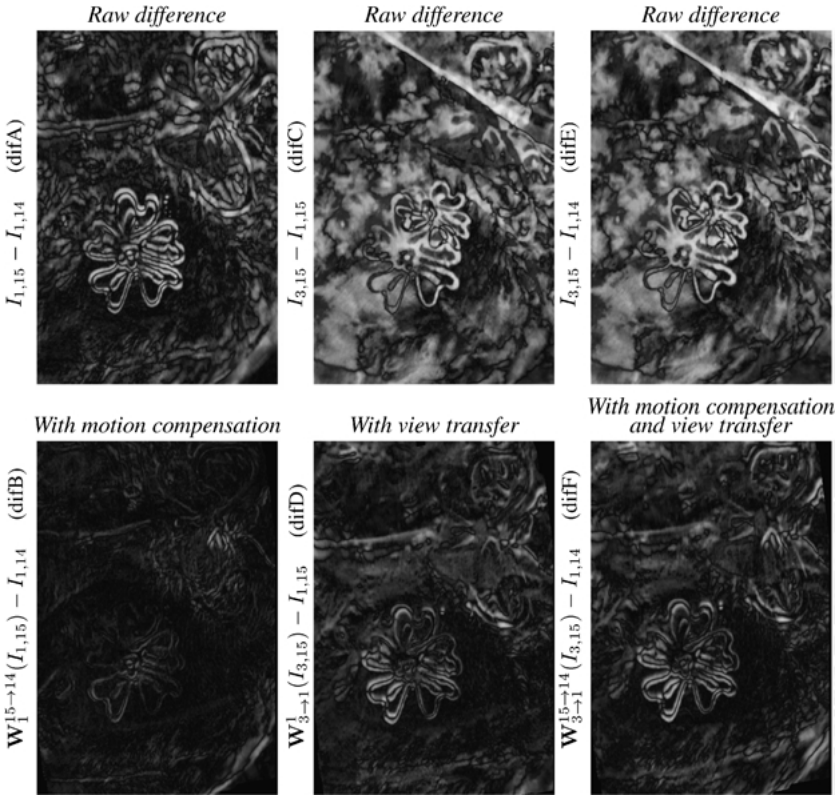

Figure 18. Intensities of pairwise differences between images in Fig. 16. The overall accuracy of our shape and motion computations is captured by the difference images in the right-most column.

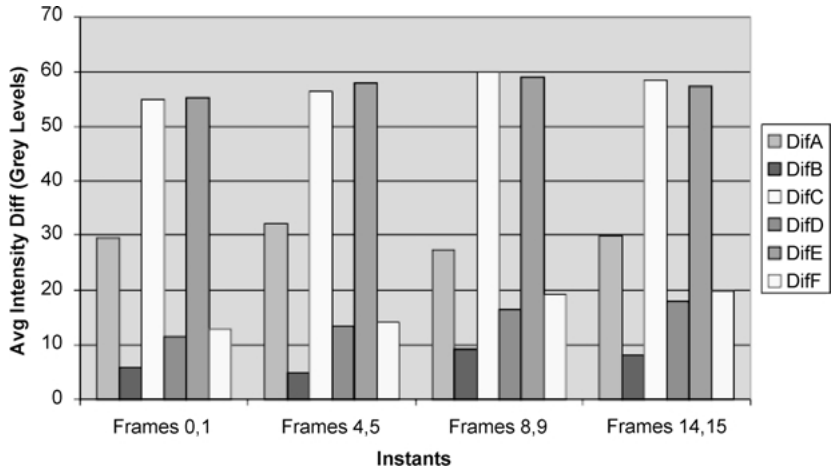

Figure 19. Average intensity values in the difference images of Figs. 17 and 18 (which correspond to Frames 4 and 14, respectively) and in the analogous difference images from Frames 0 and 8.

the shape, motion and reflectance of an unknown dynamic scene from multiple views. At the heart of this framework is the desire to explain pixels and pixel variations in the input views in terms of their underlying physical causes–shape, reflectance, motion, illumination, and visibility. We have shown that this framework leads to a shape and reflectance reconstruction algorithm called Surfel Sampling and a motion field

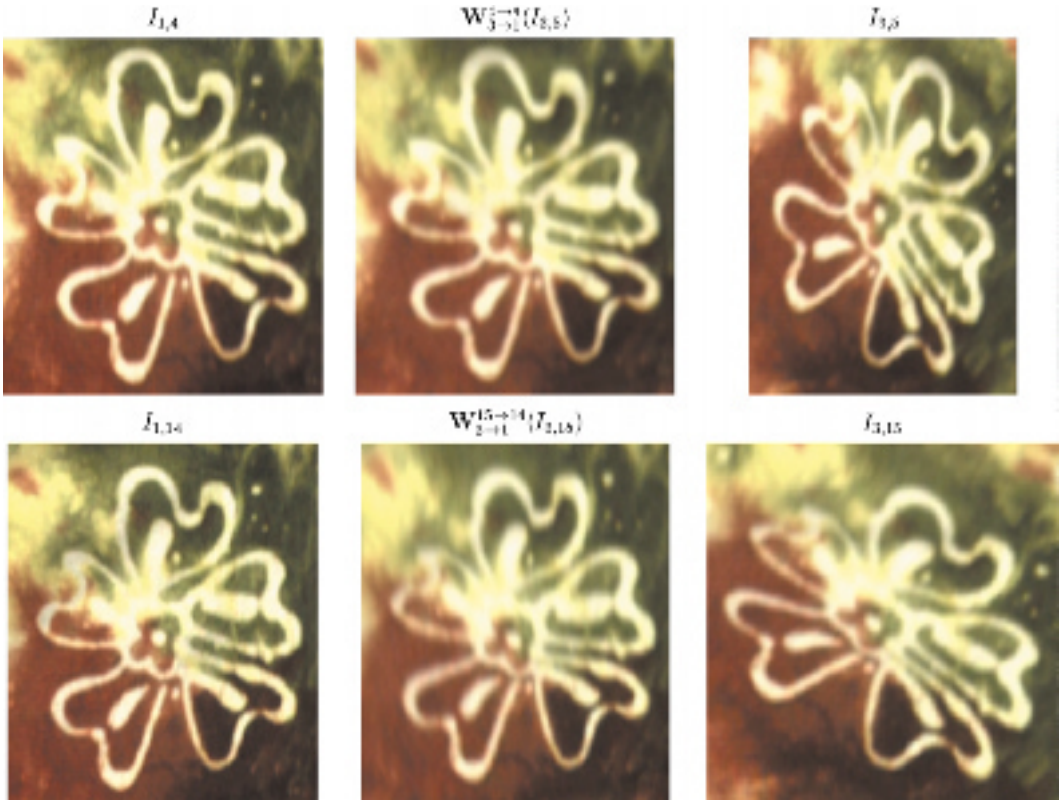

Figure 20. Enlarged views of a flower that appears prominently in the center-left portion of the images in Figs. 15 and 16.

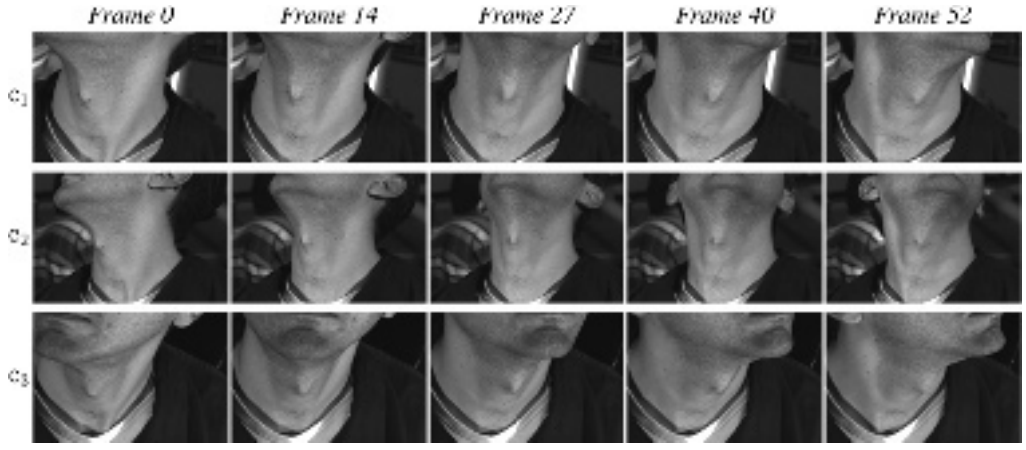

Figure 21. The "neck" sequence: a man rotates his head. Five out of ninety frames and three out of seven views are shown.

recovery algorithm called Dynamic Surfel Reconstruction that together overcome several limitations of the current state of the art. First, they provide the means to resolve the complex interactions between occlusion, parallax, shading, illumination, and deformation when analyzing the 3D shape and motion of general scenes. Second, they provide explicit information about the surface orientation of individual scene points and are able to resolve the occlusion relationships occurring between the input views of complex scenes. Third, they

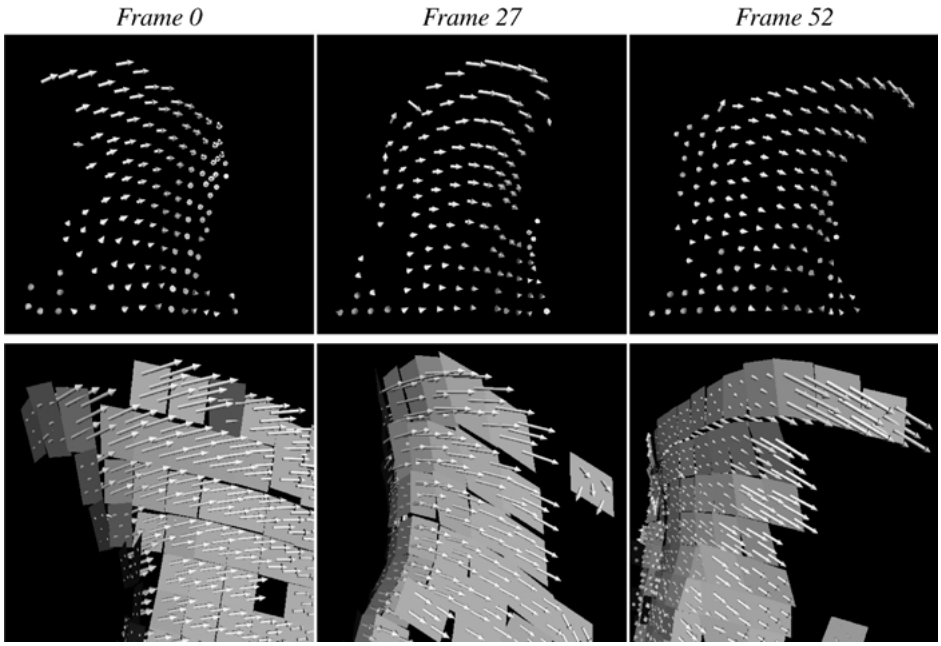

Figure 22. Reconstruction results for the "neck" sequence. *Top row*: Reconstructed 3D motion field. One vector is drawn per surfel. *Bottom row*: Close-up views of surfels and motion field in the vicinity of the chin, with five 3D motion vectors drawn per surfel.

allow us to establish explicit reconstructibility conditions that characterize their behavior for general 3D scenes. Fourth, they overcome the Lambertian reflectance constraint currently being used for 3D shape and motion recovery and can handle the presence of multiple illumination sources, shadows, and specular highlights. Fifth, they enable recovery of globally-consistent dense 3D motion fields of complex scenes without any prior information about their shape or motion.

While the effectiveness of our approach was demonstrated on a variety of complex real scenes, our framework is based on idealized models of scene illumination, reflectance, and image formation. Handling the case of unknown and possibly-extended light sources presents a formidable (and possibly intractable) inverse problem, especially when effects such as inter-reflections contribute significantly to image appearance. Despite some promising recent results that have studied this case for scenes with known geometry (Yu et al., 1999; Ramamoorthi and Hanrahan, 2001), we know of no solution to the general 3D scene capture problem under general illumination conditions. Extending our work to incorporate measurement errors and to rely on more realistic models of diffuse and specular reflectance are also topics of our current research. Other directions include (1) developing coarse-to-fine surfel sampling algorithms for capturing fine surface detail, (2) developing methods for identifying surface creases on piecewise-smooth scenes, (3) developing fast methods for non-uniform sampling of surfel-space, (4) exploiting spatio-temporal coherence while sampling surfel-space across multiple time instants, (5) studying ways to incorporate domain-dependent, spatial coherence constraints between surfels (Ju et al., 1996), and (6) investigating applications of our algorithms to image-based rendering and computer animation.

Appendix A: Surfel-Induced Image Homographies

The 3D shape component of every surfel defines a set of warp functions (Wexler and Shashua, 1999), $\mathbf{W}_{r\to i}(\,)$, that map a pixel along the "reference" view $\mathbf{c}_r$ to its corresponding pixel in the i-th view, $\mathbf{c}_i$, $i = 1, \ldots, N$. Given a surfel shape component $\mathcal{S} = \langle \mathbf{o},\ \epsilon,\ \mathbf{n},\ d \rangle$ and the 3×4 projection matrix $[\mathbf{R}_i \quad \mathbf{t}_i]$ of the i-th camera, these warp functions can be expressed as 3×3 homogeneous homography matrices (Faugeras and Keriven, 1998):

$$\mathbf{H}_i = \left(\mathbf{s}_0^T \mathbf{n}\right)\mathbf{R}_i + \mathbf{t}_i \mathbf{n}^T \tag{29}$$

$$\mathbf{q}_i = \mathbf{W}_{r\to i}(\mathbf{q}) = \mathbf{H}_i \mathbf{H}_r^{-1}\, \mathbf{q}, \tag{30}$$

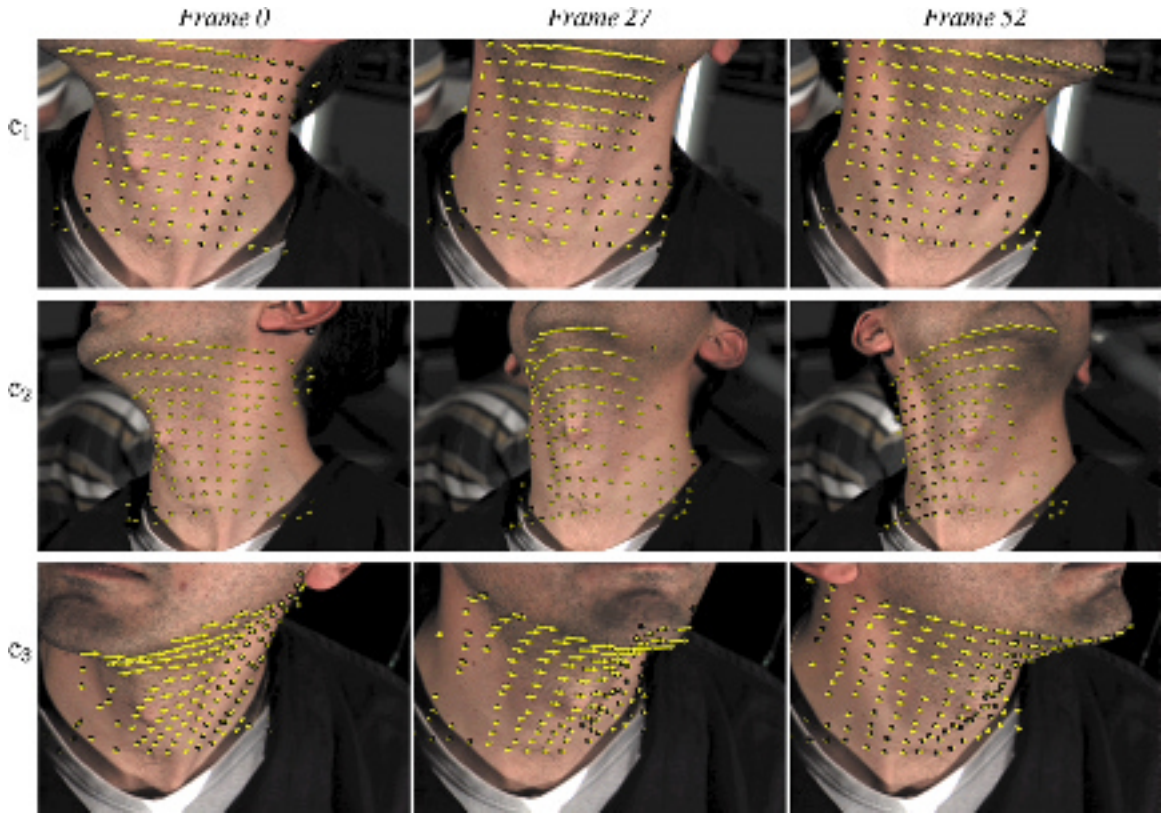

Figure 23. Reconstructed 3D motion field, viewed from the position of the input cameras and overlaid with the original "neck" sequence. Note that the left side of the neck in Frame 27 is occluded from viewpoint $\mathbf{c}_3$ but the neck's motion is correctly recovered, as indicated in the view along $\mathbf{c}_1$. Also note that since we did not perform "hidden vector" elimination when overlaying images and vectors, the vectors that are overlaid on the chin in $\mathbf{c}_3$ correspond to points that are occluded from that view.

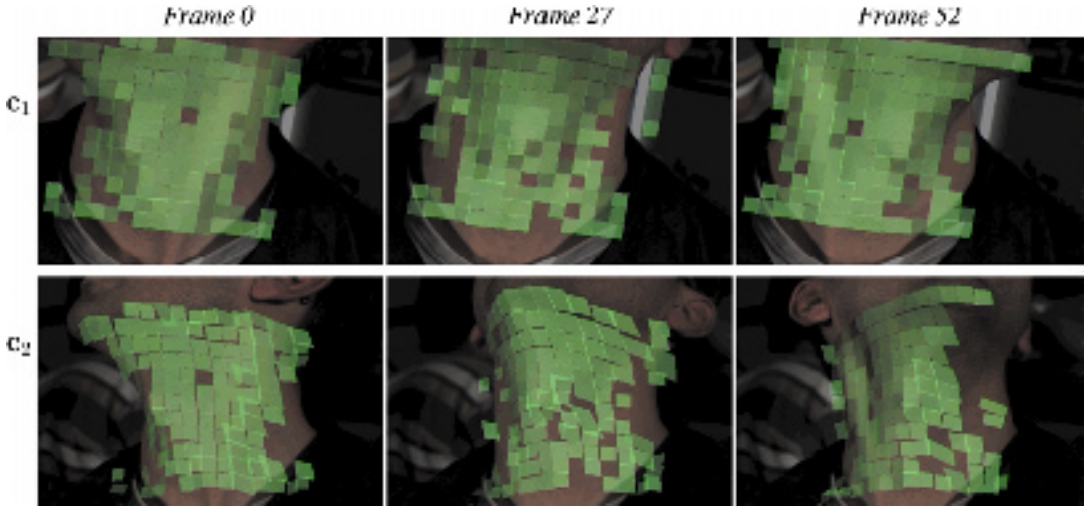

Figure 24. Shape reconstruction results for the "neck" sequence, overlaid with the input images. Reconstructed surfels are shown as transparent green regions. Since the reference camera was $\mathbf{c}_1$ for all surfels, only surface regions fully-visible to that camera were reconstructed.

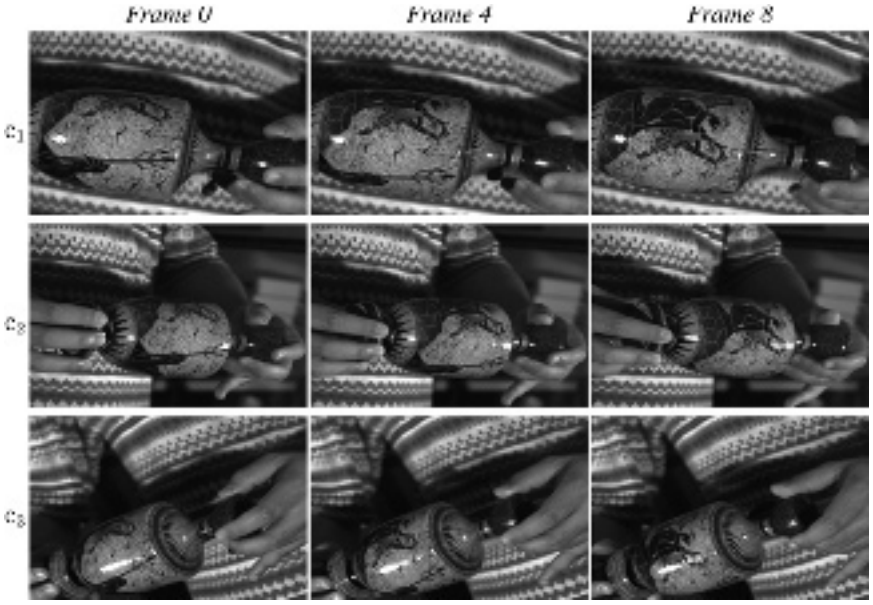

Figure 25. The "vase" sequence: a vase is manually rotated about its axis of symmetry. Three out of sixty frames and three out of seven views are shown.

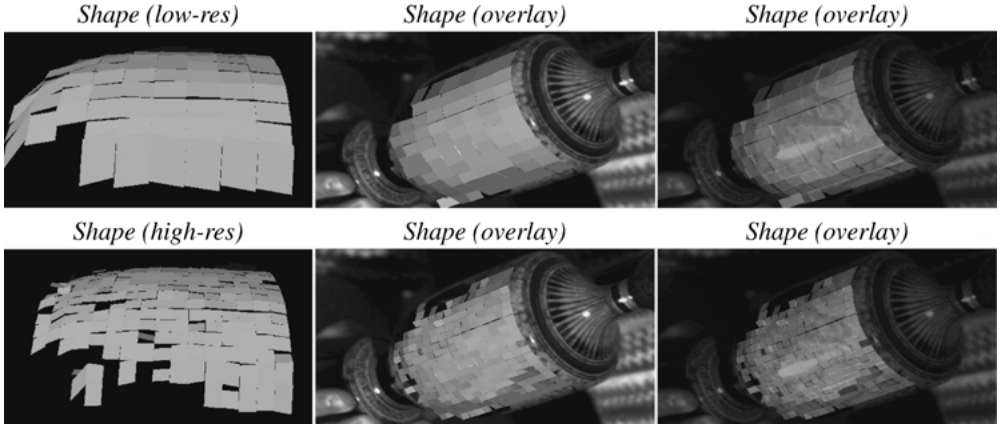

Figure 26. Shape reconstruction results for the "vase" sequence. *Top row*: Reconstructed surfels overlaid with the input images for a $16 \times 12 \times 8$-voxel tessellation of the working volume. Surfels in this volume have footprints of approximately 200 to 1000 pixels in the input views. *Bottom row*: Results for a run of the Surfel Sampling algorithm on the same images but with a higher-resolution, $32 \times 24 \times 8$ tessellation of the scene volume. In this case, surfels are one-quarter the size of surfels in the top row and have footprints that cover approximately to 50 to 250 pixels. Note that in both cases the surface regions containing specular highlights are reconstructed correctly.

where $\mathbf{s}_0$ is the surfel point closest to $\mathbf{o}$, all pixels $\mathbf{q}_i$ are expressed in homogeneous coordinates, and all equalities are up to a homogeneous scale factor.

When a surfel is allowed to move or deform, the surfel's shape and motion components define a set of warp functions $\mathbf{W}_{r\to i}^{t_0\to t}(\,)$ that map a pixel along the reference view at time t_0 to its corresponding pixel at time t in the i-th view. These warps can be expressed as compositions of three homographies and, as such, are homographies themselves:

$$\mathbf{q}_{i,t} = \mathbf{W}_{r\to i}^{t_0\to t}(\mathbf{q}) = \mathbf{H}_{i,t}\mathbf{M}_t\,\mathbf{H}_r^{-1}\,\mathbf{q}. \qquad (31)$$

In the equation above, $\mathbf{H}_{i,t}$ is obtained by replacing the static quantities $\mathbf{s}_0$ and $\mathbf{n}$ in Eq. (29) with their dynamic

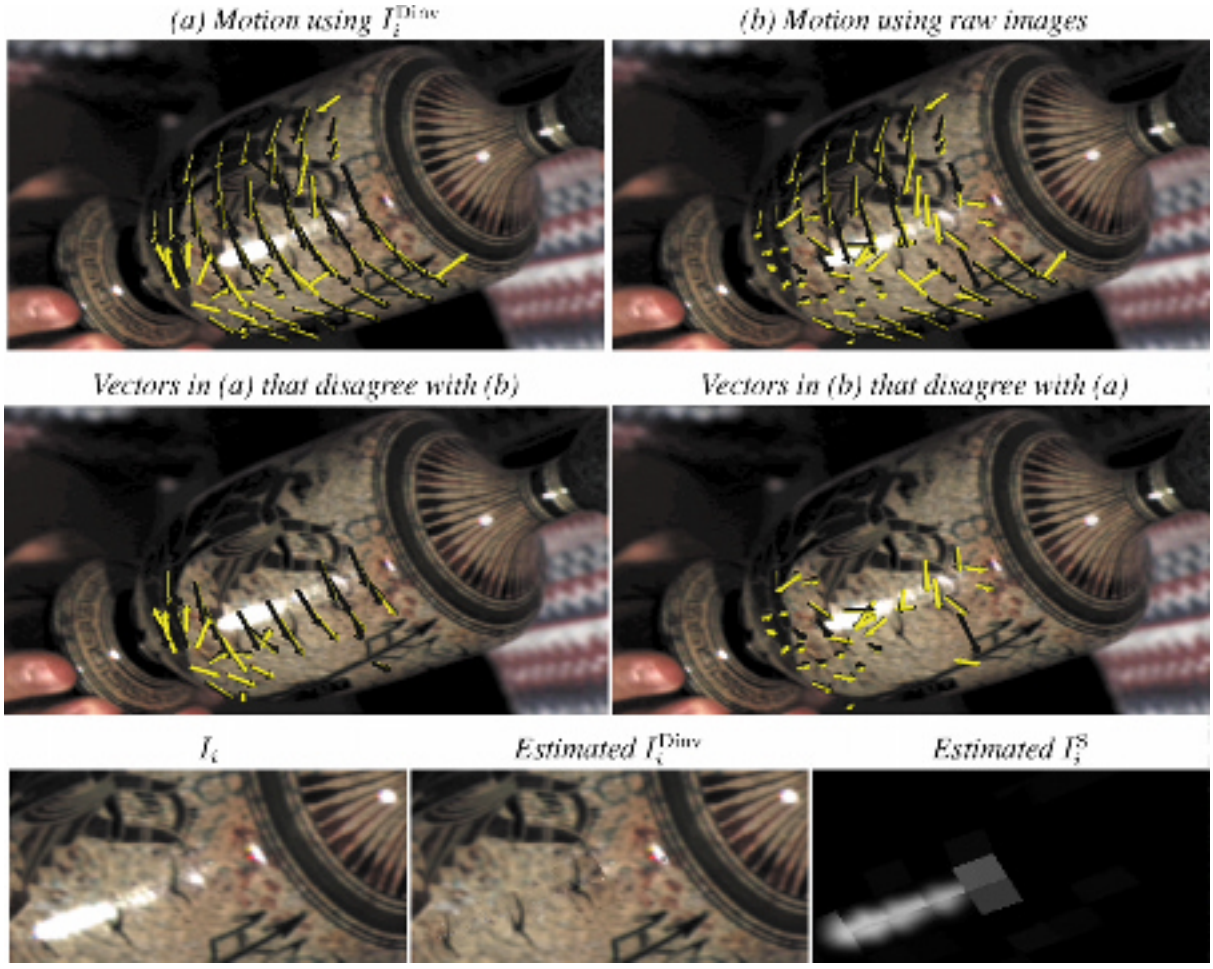

Figure 27. Reconstructed 3D motion field, overlaid with the original "vase" sequence. To illustrate the effect of our invariant-based approach, where specularities are removed from the input views before performing 3D motion computations, we show the 3D motion fields computed using our method and using an analogous method that acts directly on the raw input images (i.e., replaces I_i^{Dinv} with I_i in Eqs. (27) and (28)). *Top row*: Motion fields computed by the two methods. The black color of some vectors is due to vector rendering artifacts and is of no significance to our results. *Middle row*: Motion vectors whose counterparts in the other method's reconstruction differ significantly. Note that the approach based on raw images leads to significant errors in the motion estimates in the neighborhood of the strong specular highlight. *Bottom row*: Enlarged views of the elongated specular highlight that appears in the input images, along with its decomposition into diffuse and specular components.

counterparts, $\mathbf{s}_{0,t}$ and $\mathbf{n}_t$, respectively:

$$\begin{aligned} \mathbf{s}_{0,t} &= \mathbf{s}_0 + (t - t_0)\,\hat{\mathbf{s}}_t \quad \text{and} \\ \mathbf{n}_t &= \frac{\mathbf{n} + (t - t_0)\,(\mathbf{s}_u \wedge \hat{\mathbf{s}}_{vt} + \hat{\mathbf{s}}_{ut} \wedge \mathbf{s}_v)}{\|\mathbf{n} + (t - t_0)\,(\mathbf{s}_u \wedge \hat{\mathbf{s}}_{vt} + \hat{\mathbf{s}}_{ut} \wedge \mathbf{s}_v)\|}. \end{aligned} \tag{32}$$

The homography $\mathbf{M}_t$ induced by the surfel's 3D motion component is given by

$$\mathbf{M}_t = \left(\mathbf{s}_0^T\,\mathbf{n}\right)\left[\mathbf{I}_{3\times 3} + (t - t_0)\,\mathbf{A}\right] + (t - t_0)\,\mathbf{b}\mathbf{n}^T, \quad \text{with} \tag{33}$$

$$\mathbf{A} = \hat{\mathbf{s}}_{ut}\mathbf{s}_u^T + \hat{\mathbf{s}}_{vt}\mathbf{s}_v^T, \quad \text{and} \tag{34}$$

$$\mathbf{b} = \hat{\mathbf{s}}_t - \mathbf{A}\,\mathbf{s}_0, \tag{35}$$

where $\mathbf{I}_{3\times 3}$ is the 3×3 identity matrix.

Appendix B: Proof Sketch of Theorem 1

Without loss of generality, we prove a stronger version of Theorem 1 in which the points $\mathbf{p}$ and $\mathbf{o}$ are identical, $\mathbf{n}$ is along the normal at $\mathbf{p}$, and $d = 0$:

Theorem 1 (*Surfel Approximation Theorem*). *For every scene point* $\mathbf{p}$ *with normal* $\mathbf{n}$ *that does not project to a shadow boundary or an occlusion boundary there exists a surfel shape component* $\mathcal{S}_{\mathbf{p}}^{\epsilon} = \langle \mathbf{p}, \epsilon, \mathbf{n}, 0 \rangle$ *and components* $\mathcal{R}, \mathcal{B}$ *such that* $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B}) < \delta$.

We prove Theorem 1 by deriving a closed-form upper bound for the surfel's radius, ϵ, that satisfies $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B}) < \delta$ for an arbitrary $\delta > 0$. To obtain this bound, we analyze the magnitude of the distance on S between two scene points, $\mathbf{p}_1, \mathbf{p}_2$, that are *back-projections* of the same point on $\mathcal{S}_{\mathbf{p}}^{\epsilon}$ (Fig. 28(a)). In particular, we first show in Section B.1 that if this distance is bounded, so is $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$. We then show in Section B.2 that given an arbitrary bound on $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$, we can choose the radius ϵ so that this error is smaller than that bound. For simplicity, we restrict our analysis to the case of scenes with Lambertian reflectance; we briefly discuss how this analysis can be generalized to the case of non-Lambertian scenes at the end of Section B.2.

We use the following notation below. Given a point $\mathbf{p}_1 \in \Re^3$, $\Pi_{\mathbf{p}_1}$ is the normal plane of $\mathbf{p}$ that contains $\mathbf{p}_1$; $C_{\mathbf{p}_1}$ is the normal section $S \cap \Pi_{\mathbf{p}_1}$; and $d_M(\mathbf{p}_1, \mathbf{p}_2)$ is the distance between two 3D points $\mathbf{p}_1$ and $\mathbf{p}_2$, measured on a 1D or 2D manifold M that contains them.

B.1. Frequency-Domain Analysis

Let $\mathbf{x}(u, v)$ be a parameterization of S with $\mathbf{x}(0, 0) = \mathbf{p}$ and let $I(u, v) = I^{\text{Dinv}}(\mathbf{x}(u, v))$ be the function of S's Static Picture Invariant around $\mathbf{p}$. Since $\mathbf{p}$ does not project to a shadow boundary or an occlusion boundary, we can find a radius $\rho > 0$ such that all points in the closed ball $B(\mathbf{p}, \rho)$ are visible from the same cameras and light sources, and no point in $S \cap B(\mathbf{p}, \rho)$ projects onto a shadow or an occlusion boundary. Moreover, since the albedo of S has a finite power spectrum, the absence of visibility and illumination discontinuities in the neighborhood of $\mathbf{p}$ implies that $I(u, v)$ also has a finite power spectrum. In particular, the Fourier transform of $I(u, v)$ has non-null magnitude only at frequencies whose absolute value is smaller than some frequency upper bound and $I(u, v)$'s power has a finite value.[16] Under these conditions, Lemma 1 shows that a non-zero distance between the scene back-projections of a point on $\mathcal{S}_{\mathbf{p}}^{\epsilon}$ produces a bounded phase shift in the frequency domain, leading to a bounded value for the Static Photo-Consistency metric.[17] Let $\mathbf{p}_1, \mathbf{p}_2$ be two surface points in $B(\mathbf{p}, \rho)$ and suppose that $\gamma(s)$ is an arc-length parameterization of a surface curve $\gamma \subset B(\mathbf{p}, \rho)$ that connects $\mathbf{p}_1$ and $\mathbf{p}_2$ (Fig. 28):

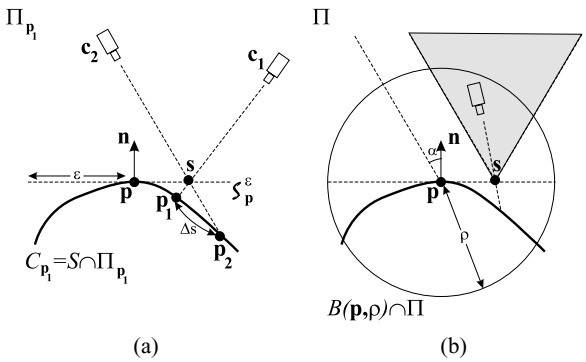

Figure 28. Geometry of Theorem 1. (a) Scene back-projections. The scene back-projections, $\mathbf{p}_1$ and $\mathbf{p}_2$, of a point $\mathbf{s} \in \Re^3$ with respect to viewpoints $\mathbf{c}_1$ and $\mathbf{c}_2$, respectively, are the visible points on the scene whose projection coincides with that of $\mathbf{s}$ for those viewpoints. (b) Given a point $\mathbf{s} \in T_{\mathbf{p}}(S) \cap B(\mathbf{p}, \rho)$, the cone whose apex is $\mathbf{s}$, whose axis is along $\mathbf{n}$, and whose generator forms an angle equal to α with $\mathbf{n}$ contains all cameras that view $\mathbf{s}$ from above $T_{\mathbf{p}}(S)$.

Lemma 1 (*Bounded Radiance Difference Lemma*). *If the Static Picture Invariant per arc length,* $I(s)$, *along every curve* γ *has all its power in the frequency range* $[-\Omega, \Omega]$, *the power of the difference* $\Delta I = I(s + \Delta s) - I(s)$ *caused by a shift* $\Delta s = d_{\gamma}(\mathbf{p}_1, \mathbf{p}_2)$ *along* γ *is bounded*:

$$|\Delta I|^2 \leq |I|^2 \, (2 - 2\cos(\min(\Omega \, \Delta s, \pi))), \qquad (36)$$

where $|I|^2$ *is the power of* $I(s)$.

Proof of Lemma 1: A positional shift in the spatial domain of a function $I(s)$ multiplies its Fourier transform, $\mathcal{I}(\omega)$, by a complex exponential factor (Pratt, 1991):

$$\mathcal{F}\{I(s + \Delta s)\} = \mathcal{I}(\omega) \, \exp(j \, \omega \, \Delta s), \quad \text{where } j \stackrel{\text{def}}{=} \sqrt{-1}.$$

Since the Fourier transform is a linear operator, the transform of the difference between the shifted and the original Static Picture Invariant is given by

$$\mathcal{F}\{I(s + \Delta s) - I(s)\} = \mathcal{I}(\omega) \, (\exp(j \, \omega \, \Delta s) - 1).$$

The integral of the squares of the terms $I(s+\Delta s)-I(s)$ over the entire curve γ, normalized by the curve's length, A, is the power of the function $I(s+\Delta s)-I(s)$. We relate this integral to the power of the function's Fourier transform, $\mathcal{F}\{I(s + \Delta s) - I(s)\}$, using

Parseval's Theorem (Pratt, 1991):

$$\begin{aligned}|\Delta I|^2 &\stackrel{\text{def}}{=} \frac{1}{A}\int_{-\infty}^{\infty} |I(s+\Delta s) - I(s)|^2\, ds \\ &= \frac{1}{2\pi A}\int_{-\infty}^{\infty} |\mathcal{F}\{I(s+\Delta s) - I(s)\}|^2\, d\omega \\ &= \frac{1}{2\pi A}\int_{-\infty}^{\infty} |\mathcal{I}(\omega)|^2 (2 - 2\cos(\omega\,\Delta s))\, d\omega \\ &\le \frac{1}{2\pi A}\int_{-\infty}^{\infty} |\mathcal{I}(\omega)|^2 \\ &\quad \times (2 - 2\cos(\min(|\omega\,\Delta s|, \pi)))\, d\omega.\end{aligned}$$

Using the hypothesis that there is a frequency upper bound, Ω, and the fact that the function $(2 - 2\cos(x))$ is monotonic for $x \in [0,\ \pi]$, the expression above yields

$$\begin{aligned}|\Delta I|^2 \le \frac{1}{2\pi A}\int_{-\infty}^{\infty} |\mathcal{I}(\omega)|^2 \\ \times (2 - 2\cos(\min(\Omega\,\Delta s, \pi)))\, d\omega.\end{aligned}$$

Since the term $(2 - 2\cos(\min(\Omega\,\Delta s, \pi)))$ does not depend on ω, it can be factored out of the integral above and Parseval's theorem can be applied again, to convert the remaining integral back to the spatial domain:

$$|\Delta I|^2 \le |I|^2\,(2 - 2\cos(\min(\Omega\,\Delta s, \pi))). \qquad \square$$

Since the Static Photo-Consistency metric, $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$, is an integral of squares of terms of the form $I(s + \Delta s) - I(s)$, normalized by the area integrated, a sufficient condition to guarantee that the measure $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$ is smaller than a given threshold δ is that the following inequalities are satisfied:[18]

$$\Omega\,\Delta s \le \pi \quad \text{and} \quad \cos(\Omega\,\Delta s) > 1 - \frac{\delta}{2|I|^2}, \qquad (37)$$

where Δs is an upper bound on the surface distance between two scene back-projections of any single point $\mathbf{s} \in \mathcal{S}_{\mathbf{p}}^{\epsilon}$. This leads to the following constraint on $\mathcal{S}_{\mathbf{p}}^{\epsilon}$:

$$\sup_{\substack{\mathbf{p}_1,\mathbf{p}_2 \in \text{backproj}(\mathbf{s}) \\ \mathbf{s} \in \mathcal{S}_{\mathbf{p}}^{\epsilon}}} d_S(\mathbf{p}_1, \mathbf{p}_2) \le \frac{1}{\Omega}\arccos\left[1 - \min\left(\frac{\delta}{2|I|^2}, 2\right)\right]. \qquad (38)$$

B.2. Spatial-Domain Analysis

We now prove a more specific version of Theorem 1, stated as follows (Fig. 28(b)):

Theorem 4 (*Surfel Approximation Theorem*)**.** *Let $\rho > 0$ be such that* (1) *all points in the closed ball $B(\mathbf{p}, \rho)$ are visible from the same cameras and light sources and* (2) *no point in $S \cap B(\mathbf{p}, \rho)$ projects onto a shadow or an occlusion boundary. The Static Photo-Consistency metric, $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$, is smaller than $\delta > 0$ if*

$$\epsilon \le (\sec\alpha - \tan\alpha)\min\left(\frac{\rho}{2},\ \frac{1}{\kappa_{max}},\ \frac{\beta}{\pi - 2\alpha}\right),$$

where β is defined by the right-hand side of Eq. (38)

$$\beta = \frac{1}{\Omega}\arccos\left[1 - \min\left(\frac{\delta}{2|I|^2}, 2\right)\right],$$

κ_{max} is the maximum absolute principal curvature of points in $S \cap B(\mathbf{p}, \rho)$; and α is the maximum angle formed by the normal $\mathbf{n}$ at $\mathbf{p}$ and a ray connecting a camera above $T_{\mathbf{p}}(S)$ to a point in $T_{\mathbf{p}}(S) \cap B(\mathbf{p}, \rho)$.

Proof of Theorem 4: We distinguish two cases:

Case A: $(\kappa_{\max} = 0)$

In this case, the surface in the ball $B(\mathbf{p}, \rho)$ is a plane and hence the tangent plane $T_{\mathbf{p}}(S)$ describes the scene's shape exactly inside this ball. It follows that the scene backprojection of every point in $T_{\mathbf{p}}(S) \cap B(\mathbf{p}, \rho)$ coincides with the point itself. Hence, $E_1(\mathcal{S}_{\mathbf{p}}^{\epsilon}, \mathcal{R}, \mathcal{B})$ will be identically zero for any $\epsilon \le \rho$.

Case B: $(\kappa_{\max} > 0)$

Let $\mathbf{s}$ be a point on $T_{\mathbf{p}}(S) \cap B(\mathbf{p}, \rho)$ and let $\mathbf{p}_1, \mathbf{p}_2$ be two scene back-projections of $\mathbf{s}$. From the triangle inequality on surfaces we have (do Carmo, 1976):

$$\begin{aligned}d_S(\mathbf{p}_1, \mathbf{p}_2) &\le d_S(\mathbf{p}_1, \mathbf{p}) + d_S(\mathbf{p}_2, \mathbf{p}) \qquad (39)\\ &\le d_{C_{\mathbf{p}_1}}(\mathbf{p}_1, \mathbf{p}) + d_{C_{\mathbf{p}_2}}(\mathbf{p}_2, \mathbf{p}). \qquad (40)\end{aligned}$$

Equation (40) suggests that we can impose a bound on the surface distance between any two scene back-projections by concentrating on the normal sections

defined by the points $\mathbf{p}_1$ and $\mathbf{p}_2$, respectively:

$$d_S(\mathbf{p}_1, \mathbf{p}_2) \leq 2 \sup_{\mathbf{p}_0} d_{C_{\mathbf{p}_0}}(\mathbf{p}_0, \mathbf{p}) \tag{41}$$

where $\mathbf{p}_0$ ranges over the set of surface points that are scene back-projections of some point in $\mathcal{S}^{\epsilon}_{\mathbf{p}}$. We now proceed by setting an upper bound for the distance $d_{C_{\mathbf{p}_0}}(\mathbf{p}_0, \mathbf{p})$ and choosing an ϵ that guarantees that no scene back-projections of points in $\mathcal{S}^{\epsilon}_{\mathbf{p}}$ can be farther away from $\mathbf{p}$. We use the following lemma for this purpose:

Lemma 2 (*Distance Bound Lemma*)**.**

$$\epsilon \leq (\sec\alpha - \tan\alpha)\rho^* \Rightarrow d_{C_{\mathbf{p}_0}}(\mathbf{p}_0, \mathbf{p}) \leq \rho^*\left(\frac{\pi}{2} - \alpha\right),$$

where

$$\rho^* = \min\left(\frac{\rho}{2}, \frac{1}{\kappa_{\max}}\right).$$

Proof of Lemma 2: Define two spheres that touch the surface at $\mathbf{p}$, have a radius equal to ρ^*, and are on opposite sides of $T_{\mathbf{p}}(S)$. By construction, these spheres are subsets of $B(\mathbf{p}, \rho)$ and every point on their surface has both principal curvatures equal to $\kappa^* = 1/\rho^*$, i.e., equal to or larger than at any point on $S \cap B(\mathbf{p}, \rho)$. It follows that the intersection of these two spheres with an arbitrary normal plane Π of $\mathbf{p}$ defines two closed disks, D_1 and D_2, bounded by circles ∂D_1 and ∂D_2, respectively, that have two properties (Fig. 29(a)): (1) D_1, D_2 are subsets of $\Pi \cap B(\mathbf{p}, \rho)$, and (2) no point of $S \cap \Pi$ is contained in the interior of D_1 and D_2.

Let $\mathbf{q}_1$ be the point on ∂D_1 whose tangent forms an angle equal to α with the normal $\mathbf{n}$ and let $\mathbf{s}$ and $\mathbf{q}_2$ be the intersections of this tangent with $T_{\mathbf{p}}(S)$ and ∂D_2, respectively (Fig. 29(b)). It follows that the scene back-projection, $\mathbf{p}_0$, of any point $\mathbf{s}_0$ on the line segment $\mathbf{sp}$ will lie inside a closed region that depends only on $\mathbf{q}_1$, D_1 and D_2:[19]

$$\mathbf{p}_0 \in \big(\overset{\triangle}{\mathbf{pq}_1\mathbf{q}_2} - D_1 \cup D_2\big). \tag{42}$$

Moreover, since the curvature of all points in $S \cap \Pi$ is equal to or smaller than that of ∂D_1 and ∂D_2, we have (do Carmo, 1976):

$$d_{C_{\mathbf{p}_0}}(\mathbf{p}_0, \mathbf{p}) \leq d_{\partial D_1}(\mathbf{q}_1, \mathbf{p}) = \rho^*\left(\frac{\pi}{2} - \alpha\right). \tag{43}$$

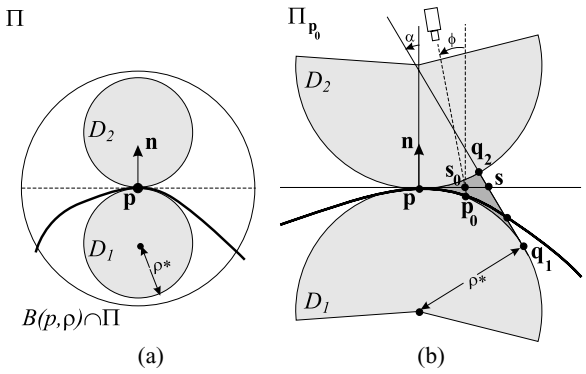

Figure 29. Proof of Lemma 2. (a) The radius ρ^* of disks D_1, D_2 is chosen so that it forces their containment inside $B(\mathbf{p}, \rho)$ and so that it is a lower bound on the radius of curvature of every normal section of every surface point in $B(\mathbf{p}, \rho)$. (b) The scene back-projection, $\mathbf{p}_0$, of every point $\mathbf{s}_0 \in \mathcal{S}^{\|\mathbf{s}-\mathbf{p}\|}_{\mathbf{p}}$ whose corresponding visual ray forms an angle $0 \leq \phi \leq \alpha$ with the surface normal will be contained in the curved-triangular region shaded in dark gray. The point $\mathbf{p}_0$ will be below $\mathbf{p}$'s tangent plane if the surface is convex at $\mathbf{p}$, will be above it if $\mathbf{p}$ is concave, and may be anywhere in the dark gray region if the surface at $\mathbf{p}$ is hyperbolic (do Carmo, 1976). Note that if $-\alpha \leq \phi \leq 0$, the region that contains $\mathbf{p}_0$ is the reflection, about the tangent plane, of the dark gray region in the figure. The case of $\phi < 0$ does not require separate treatment since it does not affect the bound on ϵ.

The lemma now follows by taking ϵ to be less than or equal to the length of segment $\mathbf{sp}$ and using the fact that $\mathbf{s}$ is the intersection of two lines that are tangent to a circle of radius ρ^* and form an angle of $(\pi/2 - \alpha)$. After some algebraic manipulations, this leads to an upper bound for ϵ that ensures satisfaction of Eq. (43):

$$\epsilon \leq (\sec\alpha - \tan\alpha)\rho^*. \tag{44}$$

□

The upper bound on ϵ in Theorem 4 now follows by choosing ϵ so that the inequalities defined by Eqs. (38), (41), (43), and (44) are satisfied simultaneously. □

A similar bound on the radius ϵ can be established for the case where the scene radiance is specular. The major difference with respect to the Lambertian case is that scene radiance is view-dependent and therefore is no longer a scalar function over the scene's surface. Nevertheless, for the specular model of Eq. (3), image intensity is still a smooth function of both the surface normal and the incoming and the outgoing directions, $\mathbf{d}^{\text{in}}$ and $\mathbf{d}^{\text{out}}$. This fact can be used to establish an upper bound on the intensity differences caused by limited errors in surface orientation, which leads to the

desired upper bound on the Static Photo-Consistency metric.

Appendix C: Proof of Theorem 2

Since only light source $\mathbf{l}_l$ contributes to the specular component of the surfel's appearance in view $\mathbf{c}_i$, the difference between the predicted and actual image intensities at the projection of surfel point $\mathbf{s}_j$ in this view can be expressed as

$$I_i^{\text{pred}}(\mathbf{s}_j) - I_i(\mathbf{s}_j) = f\,\mathcal{L}_l(\mathbf{s}_j)\big[C^{\text{S}}\big(\tilde{\mathbf{n}}_j, \mathbf{d}_j^{\text{out}}, \mathbf{d}_j^{\text{in}}\big)\big]^k - \big[I_i(\mathbf{s}_j) - I_i^{\text{Dinv}}(\mathbf{s}_j)\big], \quad (45)$$

where $I_i^{\text{Dinv}}(\mathbf{s}_j)$ is the computed static picture invariant; $\tilde{\mathbf{n}}_j$ is the bump map normal at $\mathbf{s}_j$; and $\mathbf{d}_j^{\text{out}}$ and $\mathbf{d}_j^{\text{in}}$ are the unit orientations of the rays from $\mathbf{s}_j$ to $\mathbf{c}_i$ and $\mathbf{l}_l$, respectively.

Since light source $\mathbf{l}_l$ contributes to the specular intensity at $\mathbf{s}_j$'s projection, the product $f\,\mathcal{L}_l(\mathbf{s}_j)$ in Eq. (45) cannot be zero. It follows that the left-hand side of that equation is zero for all $j = 1, \ldots, P$ if and only if for all such j

$$C^{\text{S}}\big(\tilde{\mathbf{n}}_j, \mathbf{d}_j^{\text{out}}, \mathbf{d}_j^{\text{in}}\big) - b_{ij}\, f^{-\frac{1}{k}} = 0, \quad (46)$$

with $b_{ij} = [\frac{I_i(\mathbf{s}_j) - I_i^{\text{Dinv}}(\mathbf{s}_j)}{\mathcal{L}_i(\mathbf{s}_j)}]^{\frac{1}{k}}$. Now let $\mathbf{d}_c^{\text{out}}$, $\mathbf{d}_c^{\text{in}}$ be the unit vectors along the rays $\mathbf{s}(u_c, v_c)\mathbf{c}_i$ and $\mathbf{s}(u_c, v_c)\mathbf{l}_l$, respectively. Because the distance between $\mathbf{s}_j$ and the bump map origin, $\mathbf{s}(u_c, v_c)$, is bounded by the surfel's spatial extent, the differences $\mathbf{d}_j^{\text{out}} - \mathbf{d}_c^{\text{out}}$ and $\mathbf{d}_j^{\text{in}} - \mathbf{d}_c^{\text{in}}$ both tend to zero when $\epsilon \to 0$. We can therefore re-write Eq. (46) as

$$C^{\text{S}}\big(\tilde{\mathbf{n}}_j, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}}\big) - b_{ij}\, f^{-\frac{1}{k}} = e(\epsilon) \quad (47)$$

with $\lim_{\epsilon \to 0} e(\epsilon) = 0$.

Re-writing $\tilde{\mathbf{n}}_j$ as $\mathbf{n} + \Delta\mathbf{n}_j$ in the expression above, where $\mathbf{n}$ is the surfel's normal and $\Delta\mathbf{n}_j$ collects all terms in the right-hand-side of Eq. (13) except $\mathbf{n}$, we obtain:

$$C^{\text{S}}\big(\tilde{\mathbf{n}}_j, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}}\big) = C^{\text{S}}\big(\mathbf{n}, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}}\big) + 2\big(\mathbf{d}_c^{\text{out}}\big)^T\big(\mathbf{n}\,\Delta\mathbf{n}_j^T + \Delta\mathbf{n}_j\mathbf{n}_j^T + \Delta\mathbf{n}_j\Delta\mathbf{n}_j^T\big)\big(\mathbf{d}_c^{\text{in}}\big).$$

Since the bump map's origin is taken to be the point of perfect specular reflection on the surfel, the term $C^{\text{S}}(\mathbf{n}, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}})$ is equal to one and the terms containing $\mathbf{n}\,\Delta\mathbf{n}_j^T$ and $\Delta\mathbf{n}_j\,\mathbf{n}^T$ cancel each other out, yielding:

$$C^{\text{S}}\big(\tilde{\mathbf{n}}_j, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}}\big) = 2\big(\mathbf{d}_c^{\text{out}}\big)^T\,\Delta\mathbf{n}_j\,\Delta\mathbf{n}_j^T\big(\mathbf{d}_c^{\text{in}}\big) + 1. \quad (48)$$

The theorem now follows by replacing all occurrences of $\Delta\mathbf{n}_j$ in Eq. (48) with the right-hand-side of Eq. (13) and substituting the resulting expression into Eq. (47).

Appendix D: Proof of Theorem 3

For each camera $\mathbf{c}_i$, the projection of $\mathbf{p}_j$ contributes a linear constraint to the system in Eq. (27). This constraint is given by Eq. (26), which is re-written here:

$$\left[\frac{\partial}{\partial\mathbf{p}} I_i^{\text{Dinv}}(\mathbf{p}_j)\right][\hat{\mathbf{x}}_t + u_j\,\hat{\mathbf{x}}_{ut} + v_j\,\hat{\mathbf{x}}_{vt}] - \rho_j\,\mathbf{d}^{\infty}(\mathbf{p}_j)^T\frac{\partial\mathbf{n}}{\partial t} = -\frac{\partial}{\partial t} I_i^{\text{Dinv}}(\mathbf{p}_j). \quad (49)$$

Since the moving plane can be parameterized as

$$\hat{\mathbf{x}}(u, v, t) = (\mathbf{x}_0 + u\mathbf{x}_u + v\mathbf{x}_v) + (t - t_0)\,(\hat{\mathbf{x}}_t + u\hat{\mathbf{x}}_{ut} + v\hat{\mathbf{x}}_{vt}) + \text{higher-order } (t - t_0)\text{-terms},$$

its unit normal vector at time t, $\mathbf{n}(t)$, is the unit vector in the direction of $\frac{d\hat{\mathbf{x}}}{du} \wedge \frac{d\hat{\mathbf{x}}}{dv}$:

$$\mathbf{n}(t) = \frac{\mathbf{n} + (t - t_0)(\hat{\mathbf{x}}_{ut} \wedge \mathbf{x}_v + \mathbf{x}_u \wedge \hat{\mathbf{x}}_{vt}) + \text{higher-order } (t - t_0)\text{-terms}}{\|\mathbf{n} + (t - t_0)(\hat{\mathbf{x}}_{ut} \wedge \mathbf{x}_v + \mathbf{x}_u \wedge \hat{\mathbf{x}}_{vt}) + \text{higher-order } (t - t_0)\text{-terms}\|},$$

where $\mathbf{n}$ denotes $\mathbf{n}(t_0)$. From a Taylor-series expansion of the right-hand-side above,

$$\mathbf{n}(t) = \mathbf{n} + (t - t_0)(\mathbf{I}_{3\times3} - \mathbf{n}\mathbf{n}^T)(\hat{\mathbf{x}}_{ut} \wedge \mathbf{x}_v + \mathbf{x}_u \wedge \hat{\mathbf{x}}_{vt}) + \text{higher-order } (t - t_0)\text{-terms}.$$

The vectors $\mathbf{d}_c^{\text{out}}$ and $\mathbf{d}_c^{\text{in}}$ have unit norm and hence the term $C^{\text{S}}(\tilde{\mathbf{n}}_j, \mathbf{d}_c^{\text{out}}, \mathbf{d}_c^{\text{in}})$ is equal to

$$\big(\mathbf{d}_c^{\text{out}}\big)^T\big(2\,\tilde{\mathbf{n}}_j\tilde{\mathbf{n}}_j^T - \mathbf{I}_{3\times3}\big)\big(\mathbf{d}_c^{\text{in}}\big).$$

We can therefore write the temporal derivative of $\mathbf{n}(t)$ as

$$\frac{\partial \mathbf{n}}{\partial t} = (\mathbf{I}_{3\times 3} - \mathbf{n}\mathbf{n}^T)(\hat{\mathbf{x}}_{ut} \wedge \mathbf{x}_v + \mathbf{x}_u \wedge \hat{\mathbf{x}}_{vt}) + (t - t_0)\text{-terms}. \quad (50)$$

The system in Eq. (27) now follows by substituting Eq. (50) into Eq. (49), by using the facts that $\mathbf{n} \wedge \mathbf{x}_u = \mathbf{x}_v$ and $\mathbf{x}_v \wedge \mathbf{n} = \mathbf{x}_u$, and by using $\mathbf{e}(t)$ to represent the terms of first-and-higher order with respect to $t - t_0$. By definition, the limit of $\mathbf{e}(t)$ as $t \to t_0$ is zero. □

Acknowledgments

The support of the National Science Foundation under Grant No. IIS-9875628 and of the William M. Keck Foundation under a grant for the Center for Future Health are gratefully acknowledged. Rodrigo Carceroni would also like to acknowledge the support of CNPq-Brazil, Proc. 300592/01-9.

Notes

1. While we restrict our theoretical analysis to scenes that are smooth everywhere for reasons of mathematical simplicity, our surfel representation, described in Section 3, is piecewise-smooth and our core algorithms, described in Sections 5 and 6, require only local smoothness. We therefore believe that our framework is general enough to handle piecewise-smooth scenes as well. A theoretical investigation of this topic will be the subject of future work.
2. For simplicity, we omit the color term λ in all subsequent equations.
3. In theory, every non-convex scene contains points that receive indirect illumination due to surface inter-reflections (Forsyth and Zisserman, 1991). Strictly speaking, our model is therefore only applicable to convex scenes. In practice, however, we have found this model to be quite adequate even when applied to non-convex scenes because inter-reflections usually do not dominate the image formation process for scenes that are not mirror-like.
4. We should emphasize that our approach is not dependent on the Phong model for representing reflectance. While in this work we chose the Phong model for reasons of computational simplicity, other models that better capture the reflectance properties of real scenes can also be used (e.g., the Torrance-Sparrow model (1967) and the Oren-Nayar model (1997)).
5. Note that this approach is similar in spirit to Shashua's specularity identification technique (Shashua, 1992), where specular pixels are identified in a view of the scene by estimating and subtracting the contribution of diffuse reflectance from the color of every pixel.
6. Note that $\mathbf{s}_u$ is a unit vector on the surfel's plane and hence it is determined by a single direction parameter. This vector, along with the surfel normal, uniquely define $\mathbf{s}_v$ to be the vector $\mathbf{n} \wedge \mathbf{s}_u$.
7. Even though our representation requires six parameters to represent a bump map, this representation is not minimal. This is because κ_{uv} is identically zero when the vector $\mathbf{s}_u$ is along a direction of principal surface curvature (do Carmo, 1976), resulting in a five-parameter description of the map. We avoid using this description for reasons of computational simplicity since it requires estimating the principal surface directions.
8. More formally, the theorem applies to all but a measure-zero set of scene points when the scene's surface is *generic* (Koenderink, 1990). For non-generic scenes, the set of excluded points can be a 2D region on the surface (e.g., the face of a cube viewed face-on from an input camera). Note, however, that the theorem can still be applied to points projecting to an occlusion boundary along a view $\mathbf{c}_i$ by simply excluding $\mathbf{c}_i$ from the computation of E_1. Theorem 1 can never be applied to points on a shadow boundary.
9. In this respect, our method is similar in spirit to approaches that treat specularities as outliers for motion estimation (Black et al., 2000; Black and Anandan, 1996) and recognition (Shashua, 1992). Unlike these approaches, however, our bump map estimation method goes a step further by actually using the detected specularities to extract additional information about the scene's local surface geometry.
10. While it is possible to include $\mathcal{B}$ and k in the minimization, we have found that in practice this significantly increases the number of iterations until convergence without causing a substantial reduction in the metric E_1.
11. In practice, this condition can be easily satisfied by acquiring sufficiently-dense image sequences. For instance, points moving at 3 cm/sec, observed at 30 frames/sec, and illuminated by two light sources 3 m away, induce an inter-frame 3D displacement of 1 mm/frame and an inter-frame change of less than 0.02 degrees in the orientation of $\mathbf{d}(\mathbf{p})$.
12. To form the system defined by Eqs. (27) and (28) we must compute spatial and temporal derivatives of the Static Picture Invariant, $I_i^{\text{Dinv}}(\mathbf{p}_j)$, at the projections of every sample point $\mathbf{p}_j$. In practice, we compute these derivatives at sub-pixel resolution by bilinearly interpolating their values at pixel centers. Central values are computed at each level of the Gaussian pyramid by simple differencing in space and/or time.
13. Note that this is only an approximation to $\text{vis}(\mathbf{c}_i, \mathbf{p})$ because $\mathbf{p} \in v_m$ may be occluded by points inside the voxel v_m. In practice, this implies that the size of v_m must be small enough to ensure that either such self-occlusions do not occur or that the cameras that *may* exhibit such self-occlusions can be identified without knowing the scene's shape inside v_m (Kutulakos, 2000).
14. See Kutulakos and Seitz (2000) for ways to perform such visibility computations efficiently.
15. We should note, however, that since the position of each scene point and of the cameras is known exactly, these effects *can* be accounted for with an appropriate model for the camera's lenses.
16. We define the *power* of a function $f(x)$ to be the average value of $|f(x)|^2$ over its entire domain.
17. Note that a similar analysis was used in the design of filters for optical flow estimation (Fleet and Jepson, 1990). The main difference here is that we study the relationship between displacements along a 3D surface and phase shifts on this 3D surface's radiance, instead of the relationship between image-plane displacements and phase shifts on image intensities.

18. By changing the integrals in the proof of Lemma 1 to double integrals, it is easy to show that the inequality of Lemma 1 also holds when $|\Delta I|^2$ is evaluated over a 2D neighborhood instead of the curve γ.
19. Without loss of generality, we assume here that the camera defining the scene-backprojection also lies on the plane Π. It is possible to show that the scene back-projections of $\mathbf{s}_0$ from cameras not on Π will be closer to $\mathbf{p}$ than $\mathbf{s}_0$ when measured on the surface and, hence, do not affect our distance bound. Due to space considerations, this step is omited from the proof.

References

Amenta, N., Bern, M., and Kamvysselis, M. 1998. A new Voronoi-based surface reconstruction algorithm. In *Proc. SIGGRAPH'98*, pp. 415–421.

Anandan, P. 1989. A computational framework and an algorithm for the measurement of visual motion. *Int. J. Computer Vision*, 2:283–310.

Avidan, S. and Shashua, A. 2000. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(4):348–357.

Baraff, D. and Witkin, A. 1998. Large steps in cloth simulation. In *Proc. SIGGRAPH'98*, pp. 43–54.

Belhumeur, P.N. 1996. A Bayesian approach to binocular stereopsis. *Int. J. Computer Vision*, 19(3):237–260.

Ben-Ezra, M., Peleg, S., and Werman, M. 2000. Real-time motion analysis with linear programming. *Computer Vision and Image Understanding*, 78(1):32–52.

Bérériat, D., Herlin, I., and Younes, L. 2000. A generalized optical flow constraint and its physical interpretation. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 487–492.

Black, M.J. 1999. Explaining optical flow events with parameterized spatio-temporal models. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 326–332.

Black, M.J. and Anandan, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104.

Black, M.J., Fleet, D.J., and Yacoob, Y. 2000. Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 78(1):8–31.

Blake, A. and Bulthoff, H. 1991. Shape from specularities: Computation and psychophysics. *Phil. Trans. R. Soc. Lond.*, 331:237–252.

Blinn, J.F. 1978. Simulation of wrinkled surfaces. *Computer Graphics*, 12(3):286–292.

Bouguet, J.-Y. and Perona, P. 1998. 3D photography on your desk. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 43–50.

Bregler, C., Hertzmann, A., and Biermann, H. 2000. Recovering non-rigid 3D shape from image streams. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 690–696.

Bregler, C. and Malik, J. 1998. Tracking people with twists and exponential maps. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 8–15.

Brodsky, T., Fermuller, C., and Aloimonos, Y. 1999. Shape from video. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 146–151.

Burt, P.J. and Adelson, E.H. 1983. The Laplacian pyramid as a compact image code. *IEEE Trans. on Communications*, 31(4):532–540.

Carceroni, R.L. and Kutulakos, K.N. 1999a. Toward recovering shape and motion of 3D curves from multi-view image sequences. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 192–197.

Carceroni, R.L. and Kutulakos, K.N. 1999b. Multi-view 3D shape and motion recovery on the spatio-temporal curve manifold. In *Proc. 7th Int. Conf. on Computer Vision.*, vol. 1, pp. 520–527.

Caspi, Y. and Irani, M. 2000. A step towards sequence-to-sequence alignment. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 682–689.

Chen, Q. and Medioni, G. 1999. A volumetric stereo matching method: Application to image-based modeling. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 29–34.

Collins, R.T. 1996. A space-sweep approach to true multi-image matching. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 358–363.

Cook, R. and Torrance, K.E. 1981. A reflectance model for computer graphics. *Computer Graphics*, 15:307–316.

DeCarlo, D. and Metaxas, D. 1998. Deformable model-based shape and motion analysis from images using motion residual error. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 113–119.

DeCarlo, D. and Metaxas, D. 2000. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Computer Vision*, 38(2):99–127.

Delamare, Q. and Faugeras, O. 1999. 3D articulated models and multi-view tracking with silhouettes. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 2, pp. 716–721.

Deutscher, J., Blake, A., and Reid, I. 2000. Articulated body motion capture by annealed particle filtering. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 126–133.

do Carmo, M.P. 1976. *Differential Geometry of Curves and Surfaces*. Prentice-Hall: Englewood Cliffs, NJ.

Drummond, T. and Cipolla, R. 2000. Real-time tracking of multiple articulated structures in multiple views. In *Proc. 6th European Conf. on Computer Vision*, vol. 2, pp. 20–36.

Faugeras, O. and Keriven, R. 1998. Complete dense stereovision using level set methods. In *Proc. 5th European Conf. on Computer Vision*, pp. 379–393.

Faugeras, O.D. and Keriven, R. 1998. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Trans. Image Processing*, 7(3):336–344.

Fleet, D.J., Black, M.J., Yacoob, Y., and Jepson, A.D. 2000. Design and use of linear models for image motion analysis. *Int. J. Computer Vision*, 35(3):169–191.

Fleet, D.J. and Jepson, A.D. 1990. Computation of component image velocity from local phase information. *Int. J. Computer Vision*, 5(1):77–104.

Foley, J.D., van Dam, A., Feiner, S.K., and Hughes, J.F. 1990. *Computer Graphics Principles and Practice*. Addison-Wesley.

Forsyth, D. and Zisserman, A. 1991. Reflections on shading. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(7):671–679.

Fua, P. 1997. From multiple stereo views to multiple 3-D surfaces. *Int. J. Computer Vision*, 24(1):19–35.

Fua, P. 1999. Using model-driven bundle-adjustment to model heads from raw video image sequences. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 1, pp. 46–53.

Fua, P. and Leclerc, Y.G. 1995. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int. J. Computer Vision*, 16:35–56.

Gaucher, L. and Medioni, G. 1999. Accurate motion flow estimation with discontinuities. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 2, pp. 695–702.

Guenter, B., Grimm, C., Malvar, H., and Wood, D. 1998. Making faces. In *Proc. SIGGRAPH'98*, pp. 55–66.

Haussecker, H.W. and Fleet, D.J. 2000. Computing optical flow with physical models of brightness variation. In *Proc. Computer Vision and Pattern Recogition Conf.*, vol. 2, pp. 760–767.

Horn, B.K.P. 1986. *Robot Vision*. MIT Press.

Irani, M. 1999. Multi-frame optical flow estimation using subspace constraints. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 1, pp. 626–633.

Irani, M. and Peleg, S. 1991. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53:231–239.

Irani, M., Rousso, B., and Peleg, S. 1997. Recovery of ego-motion using region alignment. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(3):268–272.

Jin, H., Yezzi, A., and Soatto, S. 2000. Integrating multi-frame shape cues in a variational framework. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 169–176.

Ju, S.X., Black, M.J., and Jepson, A.D. 1996. Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency. In *Proc. Computer Vision Pattern Recognition Conf.*, pp. 307–314.

Kanatani, K. and Ohta, N. 1999. Accuracy bounds and optimal computation of homography for image mosaicing applications. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 1, pp. 73–78.

Koenderink, J.J. 1990. *Solid Shape*. MIT Press.

Koenderink, J.J., Doorn, A.J.V., Dana, K.J., and Nayar, S. 1999. Bidirectional reflection distribution of thoroughly pitted surfaces. *Int. J. Computer Vision*, 31(2/3):129–144.

Kutulakos, K.N. 2000. Approximate N-View stereo. In *Proc. 6th European Conf. on Computer Vision*, vol. 1, pp. 67–83.

Kutulakos, K.N. and Seitz, S.M. 2000. A theory of shape by space carving. *Int. J. Computer Vision*, 38(3):199–218. Marr Prize Special Issue.

Lafortune, E.P.F., Foo, S., Torrance, K.E., and Greenberg, D.P. 1997. Non-linear approximation of reflectance functions. In *Proc. SIGGRAPH'97*, pp. 117–126.

Langer, M.S. and Zucker, S.W. 1994. Shape-from-shading on a cloudy day. *J. Opt. Soc. Am. A*, 11(2):467–478.

Lin, S. and Lee, S.W. 1999. A representation of specular appearance. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 2, pp. 849–854.

Lin, S. and Lee, S.W. 2000. An appearance representation for multiple reflection components. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 105–110.

Loop, C. and Zhang, Z. 1999. Computing rectifying homographies for stereo vision. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 125–131.

Lowe, D.G. 1991. Fitting parameterized three-dimensional models to images. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(5):441–449.

Lu, R., Koenderinck, J.J., and Cappers, A.M.L. 1999. Specularities on surfaces with tangential hairs or grooves. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 1, pp. 2–7.

Narayanan, P.J., Rander, P.W., and Kanade, T. 1998. Constructing virtual worlds using dense stereo. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 3–10.

Nayar, S.K., Fang, X., and Boult, T.E. 1993. Removal of specularities using color and polarization. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 583–590.

Negahdaripour, S. 1998. Revised definition of optical frow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(9):961–979.

Ohta, Y. and Kanade, T. 1985. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Trans. Pattern Anal. Machine Intell.*, 7(2):139–154.

Oren, M. and Nayar, S.K. 1997. A theory of specular surface geometry. *Int. J. Computer Vision*, 24(2):105–124.

Papin, C., Bouthemy, P., and Rochard, G. 2000. Tracking and characterization of highly deformable cloud structures. In *Proc. 6th European Conf. on Computer Vision*, vol. 2, pp. 428–442.

Pratt, W.K. 1991. *Digital Image Processing*. John Wiley & Sons.

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

Ramamoorthi, R. and Hanrahan, P. 2001. A signal processing framework for inverse rendering. In *Proc. SIGGRAPH'01*, pp. 117–128.

Roy, S. and Cox, I.J. 1998. A maximum-flow formulation of the N-camera stereo correspondence problem. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 492–499.

Samaras, D. and Metaxas, D. 1998. Incorporating illumination constraints in deformable models. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 322–329.

Sato, Y. and Ikeuchi, K. 1994. Temporal-color space analysis of reflection. *J. Opt. Soc. Am. A*, 11(11):2990–3002.

Sato, Y., Wheeler, M.D., and Ikeuchi, K. 1997. Object shape and reflectance modeling from observation. In *Proc. SIGGRAPH'97*, pp. 379–387.

Seitz, S.M. and Dyer, C.R. 1999. Photorealistic scene reconstruction by voxel coloring. *Int. J. Computer Vision*, 35(2):151–173.

Sidenbladh, H., Black, M.J., and Fleet, D.J. 2000. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. 6th European Conf. on Computer Vision*, vol. 2, pp. 702–718.

Silva, C. and Santos-Victor, J. 2000. Intrinsic images for dense stereo matching with occlusions. In *Proc. 6th European Conf. on Computer Vision*, vol. 1, pp. 100–114.

Shashua, A. 1992. *Geometry and photometry in 3D visual recognition*. Ph.D. Thesis, MIT.

Smith, P., Drummond, T., and Cipolla, R. 2000. Motion segmentation by tracking edge information over multiple frames. In *Proc. 6th European Conf. on Computer Vision*, vol. 2, pp. 396–410.

Snow, D., Viola, P., and Zabih, R. 2000. Exact voxel occupancy with graph cuts. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 345–352.

Szeliski, R. 1996. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, 16(2):22–30.

Szeliski, R. 1999. A multi-view approach to motion and stereo. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 157–163.

Szeliski, R., Avidan, S., and Anandan, P. 2000. Layer extraction from multiple images containing reflections and transparency. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 246–253.

Szeliski, R. and Golland, P. 1998. Stereo matching with transparency and matting. In *Proc. 6th Int. Conf. on Computer Vision*, pp. 517–524.

Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154.

Torrance, K.E. and Sparrow, E.M. 1967. Theory of off-specular reflection from roughened surfaces. *J. Opt. Soc. Am.*, 57:1105–1114.

Tzovaras, D. and Grammalidis, N. 1997. Object-based coding of stereo image sequences using joint 3-D motion/disparity compensation. *IEEE Trans. on Circuits and Systems for Video Technology*, 7(2):312–327.

Vedula, S., Baker, S., Rander, P., Collins, R., and Kanade, T. 1999. Three-dimensional scene flow. In *Proc. 7th Int. Conf. on Computer Vision*, vol. 2, pp. 722–729.

Vedula, S., Baker, S., Seitz, S., and Kanade, T. 2000. Shape and motion carving in 6D. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 592–598.

Wang, J.Y. and Adelson, E.H. 1993. Layered representation for motion analysis. In *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 361–366.

Watt, A. 2000. *3D Computer Graphics*. 3rd edn., Addison-Wesley.

Wexler, Y. and Shashua, A. 1999. Q-warping: Direct computation of quadratic reference surfaces. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 1, pp. 333–338.

Wolff, L.B., Nayar, S.K., and Oren, M. 1998. Improved diffuse reflection models for computer vision. *Int. J. Computer Vision*, 30(1):55–71.

Wood, D.N., Azuma, D.I., Aldinger, K., Curless, B., and Duchamp, T. 2000. Surface light fields for 3D photography. In *Proc. SIGGRAPH'00*, pp. 287–296.

Yacoob, Y. and Davis, L.S. 2000. Learned models for estimation of rigid and articulated human motion from stationary or moving camera. *Int. J. Computer Vision*, 36(1):5–30.

Ye, M. and Haralick, R.M. 2000. Two-stage robust optical flow estimation. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 623–628.

Yu, Y., Debevec, P., Malik, J., and Hawkins, T. 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proc. SIGGRAPH'99*, pp. 215–224.

Zelnik-Manor, L. and Irani, M. 2000. Multi-frame estimation of planar motion. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(10):1105–1116.

Zhang, Y. and Kambhamettu, C. 2000. Integrated 3D scene flow and structure recovery from multiview image sequences. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 674–681.

Zhou, L. and Kambhamettu, C. 2000. Hierarchical structure and nonrigid motion recovery from monocular views. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 752–759.

Zhou, L., Kambhamettu, C., and Goldgof, D.B. 2000. Fluid structure and motion analysis from multi-spectrum 2D cloud image sequences. In *Proc. Computer Vision and Pattern Recognition Conf.*, vol. 2, pp. 744–751.