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

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

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

Abstract

In this paper we study the problem of recovering the 3D shape, reflectance, and non-rigid motion of a dynamic 3D scene. Because these properties are completely unknown, our approach uses multiple views to build a piecewise-continuous geometric and radiometric representation of the scene's trace in space-time. Basic primitive of this representation is the dynamic surfel*, which (1) encodes the instantaneous local shape, reflectance, and motion of a small 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, skin, shiny objects) illustrate our method's ability to explain pixels and pixel variations in terms of their physical causes—shape, reflectance, motion, illumination, and visibility.*

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 [1], or deformable scenes [2]; reconstructing static Lambertian scenes [3, 4]; and recovering the reflectance of static scenes with known shape [5, 6]), our focus here is on the general case. In particular, how can we capture 3D scenes whose appearance depends on time-varying interactions between shape, reflectance, illumination, and motion? Answering this question would go a long way toward reconstructing many common real-world scenes that are beyond the current state of the art, including (1) highly-deformable and geometrically-complex surfaces whose shape, motion, self-occlusions and self-shadows change through time (e.g., clothing [7]), (2) non-Lambertian surfaces with complex shape and deformation properties (e.g., mm-scale dynamic representations of the human body), and (3) static or moving 3D objects with specular surfaces [8].

We argue that general solutions to the scene capture problem must ultimately satisfy three 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, visibility, and illumination of individual scene points (which can change dramatically), and their reflectance (which usually does not).
- **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* (surface element [9]), 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 appearance under known illumination conditions.

At the heart of our approach lies the observation that when (1) an opaque scene is viewed and illuminated in a known way, (2) its reflectance is defined parametrically, and (3) inter-reflections can be ignored, it is always 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. Here we exploit this observation by re-

The support of the National Science Foundation under Grant No. IRI-9875628, of Roche Laboratories, Inc., and of the Dermatology Foundation are gratefully acknowledged.

ducing scene capture 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 as well as their temporal variation. We show that (1) every query defines a global optimization problem in the space of all surfel descriptions, and (2) we can search this space efficiently 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. Importantly, 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. Experimental results with the Phong reflectance model [10] and complex real scenes (clothing, skin) 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 scenes. While recent scene-space stereo methods can successfully model complex static 3D scenes, they rely on discrete shape representations (e.g., lines [11], voxels [3, 12, 13], and layers [4]) 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, even though mesh-, particle- and level-set stereo methods [14–17] can, in principle, model surface orientation [17], 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.

Unlike existing methods, our surfel-based representation and our space-query formalism are spatially localized. This allows us to define a tractable optimization problem and an algorithm that has predictable performance. Importantly, because orientation is explicitly represented, our approach can handle scenes with non-Lambertian reflectance. Hence, our approach can be thought of as striking a balance between (1) the need to model orientation and maintain spatial coherence, and (2) the desire to model discontinuities and to keep the representation of distant scene points separate.

In the context of motion estimation, single-view methods have relied on known models of 3D shape [2] or motion [1] to make 3D motion estimation a well-posed problem, or have focused on improving the robustness [18, 19] and physical validity [20] of 2D motion estimation. Unfortunately, the use of known 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 [21]. Even though a small number of multi-view methods have been proposed for estimating 3D motion, their reliance on potentially noisy pointwise flow calculations and on the brightness constancy assumption [22–24] restricts them to slowly-moving Lambertian scenes, where the effects of shading and shadows on scene appearance are negligible.

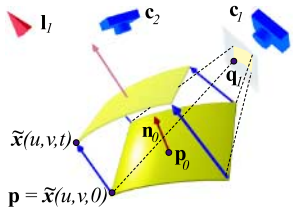

Figure 1. The dynamic surfel representation.

In this paper we argue that 3D motion estimation becomes considerably simplified when the scene's instantaneous 3D shape and reflectance properties are taken into account. 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 [25], and (5) improve shape and reflectance estimates by incorporating dynamic constraints into the surfel estimation process. In the following, we consider the scene to be an opaque, piecewise-smooth, oriented surface that is viewed under perspective projection from a collection of known viewpoints, $\mathbf{c}_1, \ldots, \mathbf{c}_N$, and that is illuminated from a set of distant point light sources at known positions, $\mathbf{l}_1, \ldots, \mathbf{l}_L$.

2. The Dynamic Surfel Representation

The dynamic surfel representation is an instantaneous description of a dynamic 3D scene in 4D space-time. Mathematically, it is a 19-degree-of-freedom analytic description of shape, reflectance, and motion, augmented by a collection of $P > 0$ discrete values that complement the surfel's reflectance description (Figure 1):

Definition 1 (Dynamic Surfel Representation). A *dynamic surfel*, $\mathcal{D}$, is represented by the tuple $\mathcal{D} = \langle \mathcal{S}, \mathcal{R}, \mathcal{M} \rangle$, where $\mathcal{S}$ is the surfel's 3D shape component; $\mathcal{R}$ is its reflectance component; and $\mathcal{M}$ is its 3D motion component.

2.1. Shape Component

Let $\mathbf{p}_0 = [x_0 \quad y_0 \quad z_0]^T$ be a smooth scene point and $\mathbf{x}$ be an orthonormal parameterization of its neighborhood. The surfel representation of $\mathbf{p}_0$'s neighborhood is simply the second-order Taylor series expansion of $\mathbf{x}$ around $\mathbf{p}_0$:

$$\mathbf{x}(u,v) = \mathbf{p}_0 + u\mathbf{x}_u + v\mathbf{x}_v + u^2\mathbf{x}_{uu} + uv\mathbf{x}_{uv} + v^2\mathbf{x}_{vv} \ ,$$

where all partial derivatives are evaluated at $\mathbf{p}_0$. The shape component of a surfel determines the surface orientation and curvature at $\mathbf{p}_0$: the direction (θ_0, ϕ_0) of its unit surface normal, $\mathbf{n}_0$, is determined by the vector product $\mathbf{x}_u \wedge \mathbf{x}_v$, while curvature is encoded by three parameters, $\kappa_1, \kappa_2, \kappa_3$, that are derived from the second-order terms of $\mathbf{x}$ [26]. This leads to the following definition for $\mathcal{S}$:

$$\mathcal{S} = \langle x_0,\ y_0,\ z_0,\ \theta_0,\ \phi_0,\ \kappa_1,\ \kappa_2,\ \kappa_3 \rangle.$$

2.2. Reflectance Component

Scene reflectance at a point $\mathbf{p}$ can be expressed as a function $\beta(\mathbf{p}, \mathbf{n}, \mathbf{d}_{\text{out}}, \mathbf{d}_{\text{in}})$ that specifies the ratio of outgoing radiance along a vector $\mathbf{d}_{\text{out}}$ to incident irradiance along a vector $\mathbf{d}_{\text{in}}$. Here we rely on the Phong reflectance model[10]:

$$\beta(\mathbf{p}, \mathbf{n}, \mathbf{d}_{\text{out}}, \mathbf{d}_{\text{in}}) = \rho(\mathbf{p}) C_{\text{L}}(\mathbf{n}, \mathbf{d}_{\text{in}}) + f\left[C_{\text{S}}(\mathbf{n}, \mathbf{d}_{\text{out}}, \mathbf{d}_{\text{in}})\right]^k \ ,$$

where $\rho(\mathbf{p})$ is the *albedo* at $\mathbf{p}$; f and k are parameters that determine the strength and extent of specular highlights; $C_{\text{L}}(\mathbf{n}, \mathbf{d}_{\text{in}})$ 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 $\mathbf{n}$. Our representation therefore models reflectance as a mixture of a diffuse component whose albedo varies from point to point, and a specular component whose coefficients are fixed for the surfel. If $\{\rho_1, \ldots, \rho_P\}$ are values of $\rho(.)$ at P surfel points, this leads to the following definition for $\mathcal{R}$:

$$\mathcal{R} = \langle f,\ k,\ \{\rho_1, \ldots, \rho_P\} \rangle.$$

When a surfel is illuminated in a known way, we can predict its appearance from $\mathcal{S}$ and $\mathcal{R}$. The pixel intensity at the projection of a surfel point $\mathbf{p}$ in camera $\mathbf{c}_i$ is given by

$$I_i^{\text{pred}}(\mathbf{p}) = \sum_{l=1}^{L} \beta(\mathbf{p}, \mathbf{n}, \mathbf{c}_i - \mathbf{p}, \mathbf{l}_l - \mathbf{p}) \mathcal{L}_l(\mathbf{p}), \qquad (1)$$

where $\mathcal{L}_l$ measures image irradiance due to the l-th light source as a function of point position. In our model, $\mathcal{L}_l(\mathbf{p})$ is zero if $\mathbf{p}$ is in shadow from $\mathbf{l}_l$ (i.e., $\mathbf{p}$ is not visible from $\mathbf{l}_l$) and is a constant $\mathcal{L}_l$ otherwise. This ignores inter-reflections and assumes that pixels measure scene radiance directly.

2.3. Motion Component

To represent dynamic surfels, we define a smooth motion field that assigns an instantaneous 3D velocity to every surfel point. Let $\tilde{\mathbf{x}}$ be a local parameterization of the surfel's spatio-temporal shape with $\tilde{\mathbf{x}}(u, v, 0) = \mathbf{x}(u, v)$. We derive the surfel's motion component in the neighborhood of $\mathbf{p}_0$ from a second-order Taylor series expansion of $\tilde{\mathbf{x}}$ that excludes acceleration terms:

$$\tilde{\mathbf{x}}(u, v, t) = \mathbf{x}(u, v) + t(\tilde{\mathbf{x}}_t + u\tilde{\mathbf{x}}_{ut} + v\tilde{\mathbf{x}}_{vt}),$$
$$\mathcal{M} = \langle \tilde{\mathbf{x}}_t,\ \tilde{\mathbf{x}}_{ut},\ \tilde{\mathbf{x}}_{vt} \rangle,$$

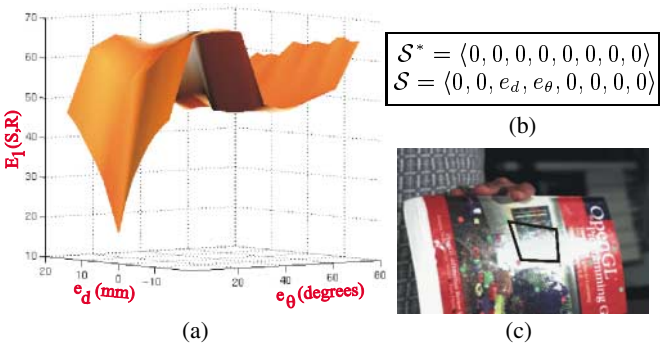

Figure 2. (a) Ground-truth plot of $E_1(\mathcal{S}, \mathcal{R})$ in the neighborhood of a planar patch, outlined in black, on the book shown in (c). (b) The true 3D position and orientation of the patch was measured in advance to obtain a ground-truth value for $\mathcal{S}^*$. E_1 is plotted with $\mathcal{S}$ defined in (b) and with $\mathcal{R}$ computed according to Section 3.2. 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 scene's strong specular properties. 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.

where all partials are evaluated at $\mathbf{p}_0$. The first term of $\mathcal{M}$ is the surfel's instantaneous translation; the remaining terms capture all motion-induced linear transformations of a plane in space. Hence, $\mathcal{M}$ can represent arbitrary translations, rotations, shearing and scaling of the surfel, but does not capture second-order deformations (i.e., changes in surface curvature). Since each partial in $\mathcal{M}$ can be an arbitrary 3D vector, $\mathcal{M}$ has nine degrees of freedom.

3. Shape & 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 $\mathcal{V}_1, \ldots, \mathcal{V}_V$ be a partitioning of $\mathcal{V}^{\text{init}}$ into V cells. To answer the above questions, we reduce global shape computation to the problem of performing a *space query* in a cell $\mathcal{V}$. In particular, suppose $\mathbf{c}$ is the position of a known camera or light source and $\omega_{\mathbf{c}}() : \mathcal{V} \rightarrow \{0, 1\}$ is a function where $\omega_{\mathbf{c}}(\mathbf{p}) = 1$ if and only if $\mathbf{p}$ is visible from $\mathbf{c}$:

Definition 2 (Space Query). Given a cell $\mathcal{V}$ and the function $\omega()$ for each camera and light source, (1) determine whether $\mathcal{V}$ contains scene points, and (2) if it does, return a surfel that is consistent with the input images and fits those points.

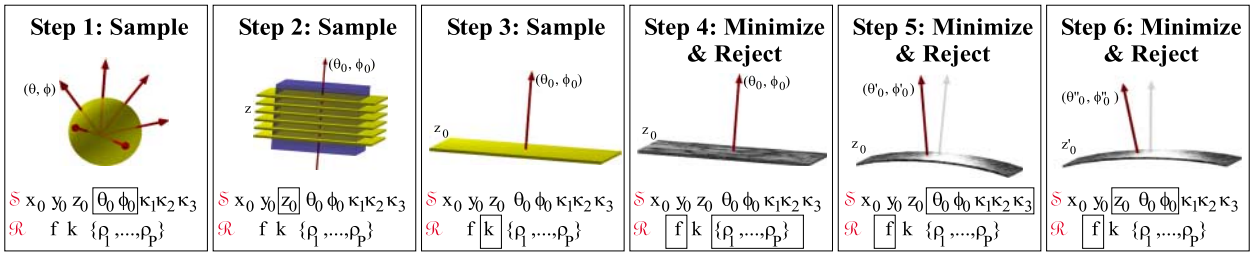

Figure 3. The Surfel Sampling Algorithm. Boxed parameters indicate the parameters to which sampling or minimization is applied. The minimization Steps 4, 5, and 6 are described in Sections 3.2, 3.3 and 3.4, respectively.

To perform a space query we must be able to determine point visibilities, i.e., know $\omega()$ for every camera and light source. Visibility determination has received considerable attention in N-view stereo research [3, 12] and is not a focus of this paper; we perform it by resolving visibilities in a way similar to the space carving algorithm [3], adapted to handle surfel-based rather than voxel-based scene representations (see Section 5 for details). In the following, we assume that all visibilities are known and concentrate on a framework we call *Surfel Sampling* which addresses the first and third questions we posed.

3.1. Space Queries by Surfel Sampling

The set of all surfels in a cell $\mathcal{V}$ is a subset of the (10+P)-dimensional space of possible shape and reflectance descriptions, $\langle \mathcal{S}, \mathcal{R} \rangle$. Surface sampling conducts an organized exploration of this space in search of the *globally optimal surfel*, $\langle \mathcal{S}^*, \mathcal{R}^* \rangle$, i.e., the surfel in $\mathcal{V}$ whose appearance globally minimizes an appropriately-defined image consistency error metric. Once this surfel is identified, its consistency with the input images is used to either reject it (i.e., set $\mathcal{V}$ empty), or accept it as a valid local scene description.

Our consistency metric measures the difference between predicted and actual pixel intensities:

$$E_1(\mathcal{S},\mathcal{R}) = \frac{1}{A_1}\sum_{i=1}^{N}\sum_{j=1}^{P}\omega_{\mathbf{c}_i}(\mathbf{p}_j)\left[I_i^{\text{pred}}(\mathbf{p}_j) - I_i(\mathbf{p}_j)\right]^2,$$

$$\langle \mathcal{S}^*, \mathcal{R}^* \rangle = \arg\min_{\mathcal{S}\in\mathcal{V}}\left[\min_{\mathcal{R}} E_1(\mathcal{S},\mathcal{R})\right],$$

where A_1 is a normalizing factor and $I_i^{\text{pred}}(\mathbf{p}_j)$ is given by Eq. (1) with $\mathcal{L}_l(\mathbf{p}) = \omega_{\mathbf{l}_l}(\mathbf{p})\mathcal{L}_l$. Given this metric, a cell is considered empty when $E_1(\mathcal{S}^*,\mathcal{R}^*)$ is greater than a fixed variance threshold σ.

Querying space by minimizing E_1 is difficult for three reasons. First, a local minimization of E_1 is not sufficient to decide if $\mathcal{V}$ 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 (Figure 2). Intuitively, this is because even a small deviation from a surfel's optimal position and orientation will corrupt stereo correspondences and specularity computations, making it hard to explain the input images. Second, the metric has deep local minima for real scenes, making it impossible to guarantee global minimization using standard methods (e.g., Levenberg-Marquardt). Third, an exhaustive search of $\langle \mathcal{S}, \mathcal{R} \rangle$-space is practically impossible because of its high dimensionality.

To overcome these difficulties, the Surfel Sampling Algorithm combines a coarse sampling of $\langle \mathcal{S}, \mathcal{R} \rangle$-space along specific dimensions with a sequence of linear and non-linear optimization steps that "explore" the neighborhood of each $\langle \mathcal{S}, \mathcal{R} \rangle$-sample. A graphical depiction of the algorithm is shown in Figure 3. 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.

3.2. Linear Reflectance Estimation

Steps 1-4 of the Surfel Sampling Algorithm are directed toward finding samples in surfel space that may be near the globally optimal surfel, $\langle \mathcal{S}^*, \mathcal{R}^* \rangle$, in $\mathcal{V}$. This is done by (1) generating a sample $\mathcal{S}$ of linear shape parameters and of the specular exponent, k, (2) augmenting $\mathcal{S}$ with an optimal assignment of linear reflectance parameters, $\mathcal{R}$, and (3) assessing the consistency of $\langle \mathcal{S}, \mathcal{R} \rangle$ with the input views.

More specifically, we observe that the image formation model of Eq. (1) becomes linear when the specular exponent, k, is known for a planar surfel with known shape, $\mathcal{S} = \langle x_0,\ y_0,\ z_0,\ \theta_0,\ \phi_0,\ 0,\ 0,\ 0 \rangle$. If $\mathbf{p}$ is a surfel point that is visible to the i-th camera, we can maximize the point's consistency with the images by solving a linear equation with only two unknowns, $\rho(\mathbf{p})$ and f. This equation forces agreement between $\mathbf{p}$'s actual and predicted intensities:

$$I_i(\mathbf{p}) = A\rho(\mathbf{p}) + Bf, \tag{2}$$

where $I_i(\mathbf{p})$ is the measured intensity at $\mathbf{p}$'s projection and A, B collect the known terms of Eq. (1). For P surfel points projecting to a total of M pixels, Eq. (2) gives rise to a linear system of M equations and $P+1$ unknowns, corresponding to the P individual albedo values and the common specular coefficient. In practice, we form the system by uniformly sampling the surfel's parameterization, $\mathbf{x}(u,v)$, in the inte-

rior of cell $\mathcal{V}$. This results in a sparse linear system that is solvable in $O(M)$ steps.

Since $\mathcal{S}$ and k must be known to estimate reflectance, we generate them through sampling. We first uniformly sample the (θ_0, ϕ_0)-space so that neighboring normals form an angle smaller than a constant ψ. Then we uniformly sample the 1D space of depths, z_0, along $\mathbf{n}_0$. These two sampling steps span the entire space of planes in $\mathcal{V}$ (Steps 1 and 2 in Figure 3). After coarsely sampling the space of k-values and estimating $\mathcal{R}$, the $\langle \mathcal{S}, \mathcal{R} \rangle$-sample is accepted for further analysis if $E_1(\mathcal{S}, \mathcal{R})$ is smaller than the threshold σ.

3.3. Curvature Compensation

Linear reflectance estimation assumes that scene points in $\mathcal{V}$ can be approximated by a plane. Ignoring curvature when the scene in $\mathcal{V}$ is curved will always generate a sub-optimal surfel solution. Importantly, it may lead to an incorrect rejection of a planar sample $\langle \mathcal{S}, \mathcal{R} \rangle$ even though it is tangent to a scene point in $\mathcal{V}$ (Figure 4a). We overcome these difficulties by noting that since we know the orientation of a surfel and the light source positions, we can reason directly about the presence or absence of specularities and about how, by interacting with surface curvature, they affect surfel appearance. We use this idea in two ways. First, we slightly modify our linear reflectance estimation step to make surfel rejection robust to curvature-induced effects. This is accomplished by estimating a surfel's reflectance from a subset of its input views, i.e., those images where curvature-induced effects due to a strong specular highlight will not be present at the surfel's projection. Second, we refine the shape component of every surfel that projects to a strong specular highlight in at least one input view (Figure 4b). Details can be found in [27].

3.4. Non-Linear Shape & Reflectance Estimation

Curvature-compensated adjustments to a surfel's position and orientation occur only when the surfel projects to a strong specular highlight. Since the linear reflectance estimation of Section 3.2 does not adjust these parameters either, the accuracy of position and orientation estimates for all other surfels is given by the density of samples in (θ_0, ϕ_0, z_0)-space. This produces sub-optimal surfel solutions. To optimize these solutions without densely sampling the (θ_0, ϕ_0, z_0)-space, we apply non-linear minimization to a metric that approximates E_1 and is independent of a surfel's P albedo parameters. Our formulation relies on a generalization of the brightness constancy constraint [28]:

Observation 1 (Generalized Brightness Constancy). If $\mathbf{p}$ is a static scene point that follows the Phong model, the intensity of its projection, after subtracting the contribution of specular reflectance, will be identical in all input views:

$$I_i^{\text{diff}}(\mathbf{p}) = I_i(\mathbf{p}) - \sum_{l=1}^{L} f\,[C_{\text{S}}(\mathbf{n}, \mathbf{c}_i - \mathbf{p}, \mathbf{l}_l - \mathbf{p})]^k = \text{const.} \quad (3)$$

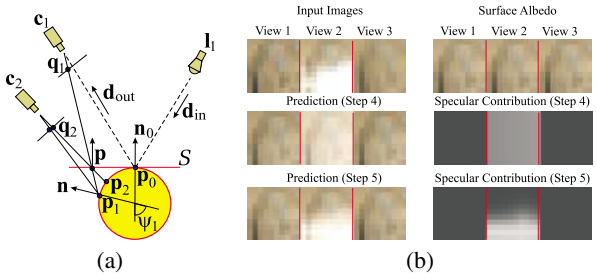

Figure 4. Curvature compensation. (a) Errors due to planar approximations of a curved scene. The normal of the scene point projecting to $\mathbf{q}_1$ is $\mathbf{n}$, not $\mathbf{n}_0$. For the specular viewpoint $\mathbf{c}_1$, the error in Eq. (1) is dominated by the error in $\mathbf{n}$—if $\mathbf{p}_1$ has a large specular exponent k, $\mathbf{q}_1$ will *not* exhibit the specularity predicted by the planar surfel assumption. To improve the robustness of linear reflectance estimation (Step 4 of the Surfel Sampling Algorithm), we ignore in that step near-specular views, i.e., those where $\arccos[C_{\text{S}}(\mathbf{n}, \mathbf{d}_{\text{out}}, \mathbf{d}_{\text{in}})] < \psi_{\text{min}}$ for at least one surfel point and one light source. The angle ψ_{min} is determined from the surfel's (known) specular exponent. (b) Reflectance estimation results for the "vase" scene in Figure 5. Shown are closeup views of a surfel (left column) and the predicted contributions to these views of each of the surfel's reflectance components (right column). Note the significant prediction errors in Step 4, where planarity is assumed.

We use this observation to reformulate our original consistency metric, E_1, into a new albedo-independent metric that is minimized by applying Levenberg-Marquardt's method to the "diffuse" images, I_i^{diff} (Step 6 in Figure 3):

$$E_2(\mathcal{S}, \mathcal{R}) = \frac{1}{A_2} \sum_{i=2}^{N} \omega_{\mathbf{c}_1}(\mathbf{p})\, \omega_{\mathbf{c}_i}(\mathbf{p}) \int_{\mathbf{p}} \left[I_1^{\text{diff}}(\mathbf{p}) - I_i^{\text{diff}}(\mathbf{p}) \right]^2 d\mathbf{p},$$

where $\mathbf{p}$ ranges over the surfel's points in $\mathcal{V}$, and A_2 is a normalizing factor. Since every surfel defines a warp function that maps a point in the reference image to its corresponding point in the other views, E_2 can be thought of as evaluating the difference between the diffuse reference image and its warped counterparts. The metric is therefore similar in spirit to warping-based metrics proposed for 2D image registration and motion estimation [4].

4. Motion Estimation By Surfel Sampling

Any general approach to the dynamic scene capture problem 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 change 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 actually be resolved when an initial 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 scene's 3D motion—they strongly constrain the scene's instantaneous 3D shape as well [23, 29]. We apply these two principles for dynamic surfel recovery by (1) developing a new linear method that estimates a surfel's motion component, $\mathcal{M}$, directly from $\langle \mathcal{S}, \mathcal{R} \rangle$ and the pixel intensities at its projection, and (2) using an additional, non-linear estimation step that jointly refines $\mathcal{S}, \mathcal{R}$ and $\mathcal{M}$ using Levenberg-Marquardt optimization to maximize their consistency with the input images. Below, we focus on our linear 3D motion estimation method. Joint non-linear minimization of $\mathcal{S}, \mathcal{R}$, and $\mathcal{M}$ follows as a straightforward, temporal generalization of our analysis in Section 3.4 and is described in [27].

4.1. Direct & Linear 3D Motion Estimation

Our linear motion estimation approach is based on a temporal analogue of the Generalized Brightness Constancy constraint. Let $\mathbf{p} = \tilde{\mathbf{x}}(u, v, t)$ be a point at time t whose visibility with respect to the i-th camera and all light sources remains unchanged near t, and suppose that light sources are distant, i.e., $\frac{d}{dt}\|\mathbf{l}_l - \mathbf{p}\| \ll \|\mathbf{l}_l - \mathbf{p}\|$. Observation 2 tells us that if we subtract the contribution of specular reflections from the input views, the intensity at $\mathbf{p}$'s projection will be constant unless $\mathbf{p}$'s orientation changes:

Observation 2 (Generalized Temporal Brightness Constancy). If $\mathbf{p}$'s albedo remains constant, the total time derivative of the diffuse image intensity, I_i^{diff}, at $\mathbf{p}$'s projection satisfies

$$\frac{d}{dt} I_i^{\text{diff}} = \rho(\mathbf{p})\, \mathbf{d}(\mathbf{p})^T \frac{\partial \mathbf{n}}{\partial t}, \tag{4}$$

with all functions evaluated at point $\mathbf{p}$, the diffuse intensity I_i^{diff} given by Eq. (3), and $\mathbf{d}(\mathbf{p})$ given by

$$\mathbf{d}(\mathbf{p}) = \sum_{l=1}^{L} \frac{\mathcal{L}_l(\mathbf{p})}{\|\mathbf{l}_l - \mathbf{p}\|} (\mathbf{l}_l - \mathbf{p}). \tag{5}$$

A key feature of this temporal constraint is that it leads to a linear system for computing a surfel's motion component from its shape and reflectance parameters and from the pixel intensities at its projection. Specifically, suppose we are given a set of points on a surfel with central point $\mathbf{p}_0$. If these points project to a total of M pixels, we have:

Theorem 1 (Linear 3D Motion Estimation Theorem). The motion $\mathcal{M}$ of a planar scene patch with known shape $\mathcal{S}$ and reflectance $\mathcal{R}$ satisfies the $M \times 9$ system

$$\mathbf{A} \begin{bmatrix} \tilde{\mathbf{x}}_t^T & \tilde{\mathbf{x}}_{ut}^T & \tilde{\mathbf{x}}_{vt}^T \end{bmatrix}^T = \mathbf{b} + \mathbf{e}(t), \tag{6}$$

where all partials are evaluated at $\mathbf{p}_0$; the terms $\mathbf{A}, \mathbf{b}$ depend only on $\langle \mathcal{S}, \mathcal{R} \rangle$ and the image intensities at the surfel's projection; and $\mathbf{e}(t)$ depends on the terms of $\frac{\partial \mathbf{n}}{\partial t}$ that are of first-or-higher-order with respect to t and, therefore, $\lim_{t \to 0} \mathbf{e}(t) = 0$.

Proof sketch. Consider a scene point $\mathbf{p} = \tilde{\mathbf{x}}(u, v, t)$ projecting to a pixel in image i. That pixel contributes a single linear equation to the system. To form the equation, we express the total derivative of pixel intensities at $\mathbf{p}$'s projection in terms of $\mathcal{M}$

$$\frac{d}{dt} I_i^{\text{diff}} = \left[\frac{\partial}{\partial \mathbf{p}} I_i^{\text{diff}} \right] [\tilde{\mathbf{x}}_t + u\tilde{\mathbf{x}}_{ut} + v\tilde{\mathbf{x}}_{vt}] + \frac{\partial}{\partial t} I_i^{\text{diff}} \tag{7}$$

and equate the right hand sides of Eqs. (4) and (7). Eq. (6) follows by expressing $\frac{\partial \mathbf{n}}{\partial t}$ at $\mathbf{p}$ in terms of the partials $\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{ut}, \tilde{\mathbf{x}}_{vt}$, evaluated at $\mathbf{p}_0$, and assigning to vector $\mathbf{e}(t)$ all terms obtained from $\frac{\partial \mathbf{n}}{\partial t}$ that are first-or-higher-order with respect to t. □

See [27] for a detailed proof. Solving Eq. (6) as a linear system for an infinitesimal time interval can be thought of as minimizing an error metric equivalent to E_1 (Section 3.1). Besides providing 3D motion estimates, Theorem 1 allows us to determine, through Singular Value Decomposition, which of $\mathcal{M}$'s parameters can be recovered reliably. This is especially important since real scenes frequently contain regions where motion information cannot be extracted because of the aperture problem [28]. In practice, Theorem 1 leads to a direct coarse-to-fine algorithm for computing $\mathcal{M}$ that consists of five steps: (1) create diffuse sequences from the input videos (Section 3.4), (2) build their Gaussian pyramids, (3) use SVD analysis to evaluate which, if any, of $\mathcal{M}$'s parameters can be reliably estimated, (4) solve the linear system of Eq. (6) for those parameters using P surfel points at level h of the pyramid, and (5) refine $\mathcal{M}$ by repeating these steps for level $h - 1$.

5. 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 TMC-9700 color cameras. The cameras allowed simultaneous observation of an approximately 30x30x30cm working volume. A sequence of geometric and radiometric calibration steps ensured that (1) the projection of points within the working volume was accurate to approximately 0.5 pixels, and (2) 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 whose 3D positions were recovered by adapting the method in [30]. We used color images for all computations, treating each band as an additional image.

The working volume was divided equally into a 16x12x8 array of cells for performing space queries, with every cell having at most one surfel that describes it. To resolve occlusions between cameras, light sources and surfel points, we used an adaptation of the Space Carving Algorithm [3]: space queries were first performed on the "surface" of the cell array and cells that were determined to be empty were "carved away" from the array. Surfel points were considered visible from a camera if they were not occluded by a surfel or by a cell whose occupancy was not yet established.

This ensured that visibilities were always estimated conservatively and that surfels could be partially visible. Shadows from one or more light sources were resolved by treating them as cameras in the above algorithm.

Figure 5 shows results of applying our Surfel Sampling algorithm to three complex scenes of significantly different shape, reflectance, and motion dynamics: (1) the "flag" sequence was created by rapidly shaking a thin near-Lambertian flag with both highly-textured and sparsely-textured regions—this resulted in dramatic 3D shape changes that caused self-occlusions and self-shadows, and induced a motion field that varied wildly over the surface in both magnitude and direction; (2) the "neck" sequence depicts neck deformations due to a head rotation and requires handling a scene with non-Lambertian reflectance, little texture, and 3D motion that causes self-occlusions and complex deformations; and (3) the "vase" sequence was created by manually rotating a highly-specular vase about its axis of symmetry, inducing a rigid 3D motion field.

Only two parameters were changed across sequences—minor adjustments to the working volume and adjustments to the error threshold for surfel rejection. We relied on the Phong model for all sequences (Section 2.2) and used exactly the same rates for sampling surfel space. Specifically, orientation sampling in Step 1 of the algorithm was achieved by choosing 193 uniformly-spaced samples of the Gaussian hemisphere facing the cameras. This corresponded to approximately 10 degrees between neighboring samples for the surfel normal. The distance between z_0-samples was 0.5mm in Step 2, and the same four values (1.5, 6, 25 and 100) were used for the specular exponent, k.

Three observations can be made from our experiments. First, despite the low sampling rate and the low resolution of the cell array, it is possible to obtain globally coherent scene descriptions *without using explicit models, global regularization constraints or smoothness biases*—in effect, the scene capture problem has sufficient "internal constraints" from geometry, reflectance, and motion, to eliminate the need for such *a priori* constraints. Second, our approach can handle scenes with significant occlusions and strong specular highlights, both of which present major challenges to common stereo algorithms. Third, our method can be used to capture, at sub-surfel resolution, the highly-complex 3D motion fields induced by surfaces such as clothing and skin that deform, self-occlude, and move in ways that vary considerably from point to point. Capturing dense fields for such surfaces is difficult to achieve through instrumentation (e.g., motion sensors) since it would require covering the surfaces in question, affecting appearance and, possibly, their mass and dynamics. We are not aware of techniques that can capture the deformation of such scenes reliably.

6. Concluding Remarks

While the effectiveness of our approach was demonstrated on a variety of scenes, our formulation relies on idealized models of reflectance and image formation. Extending this framework to incorporate measurement errors and more realistic models of diffuse and specular reflectance are topics of current research. Other directions include (1) developing fast methods for non-uniform sampling of surfel-space, (2) developing sampling methods that exploit spatio-temporal coherence, and (3) investigating applications in image-based rendering and computer animation.

References

[1] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. CVPR*, pp. 690–696, 2000.

[2] B. Guenter, C. Grimm, H. Malvar, and D. Wood, "Making faces," in *Proc. SIGGRAPH'98*, 1998. 55–66.

[3] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, v. 38, n. 3, pp. 199–218, 2000.

[4] R. Szeliski, "A multi-view approach to motion and stereo," in *Proc. CVPR*, v. 1, pp. 157–163, 1999.

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

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

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

[8] L. B. Wolff, S. K. Nayar, and M. Oren, "Improved diffuse reflection models for computer vision," *IJCV*, v. 30, n. 1, pp. 55–71, 1998.

[9] H. Pfister, M. Zwicker, J. van Baar, and M. Gross, "Surfels: Surface elements as rendering primitives," in *Proc. SIGGRAPH'00*, pp. 335–342, 2000.

[10] A. Watt, *3D Computer Graphics*. Addison-Wesley, 2000.

[11] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. CVPR*, pp. 358–363, 1996.

[12] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *IJCV*, v. 35, n. 2, pp. 151–173, 1999.

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

[14] P. Fua and Y. G. Leclerc, "Object-centered surface reconstruction: Combining multi-image stereo and shading," *IJCV*, v. 16, pp. 35–56, 1995.

[15] P. Fua, "From multiple stereo views to multiple 3-D surfaces," *IJCV*, v. 24, n. 1, pp. 19–35, 1997.

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

[17] D. Samaras and D. Metaxas, "Incorporating illumination constraints in deformable models," in *Proc. CVPR Conf.*, pp. 322–329, 1998.

[18] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *IJCV*, v. 2, pp. 283–310, 1989.

[19] M. J. Black, D. J. Fleet, and Y. Yacoob, "Robustly estimating changes in image appearance," *CVIU*, v. 78, n. 1, pp. 8–31, 2000.

[20] S. Negahdaripour, "Revised definition of optical frow: Integration of radiometric and geometric cues for dynamic scene analysis," *IEEE T-PAMI*, v. 20, n. 9, pp. 961–979, 1998.

[21] H. W. Haussecker and D. J. Fleet, "Computing optical flow with physical models of brightness variation," in *Proc. CVPR*, v. 2, pp. 760–767, 2000.

[22] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *Proc. 7th ICCV*, pp. 722–729, 1999.

[23] S. Vedula, S. Baker, S. Seitz, and T. Kanade, "Shape and motion carving in 6D," in *Proc. CVPR*, v. 2, pp. 592–598, 2000.

[24] Y. Zhang and C. Kambhamettu, "Integrated 3D scene flow and structure recovery from multiview image sequences," in *Proc. CVPR*, v. 2, pp. 674–681, 2000.

[25] M. Irani, "Multi-frame optical flow estimation using subspace constraints," in *Proc. 7th ICCV*, pp. 626–633, 1999.

[26] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*. Prentice-Hall Inc., 1976.

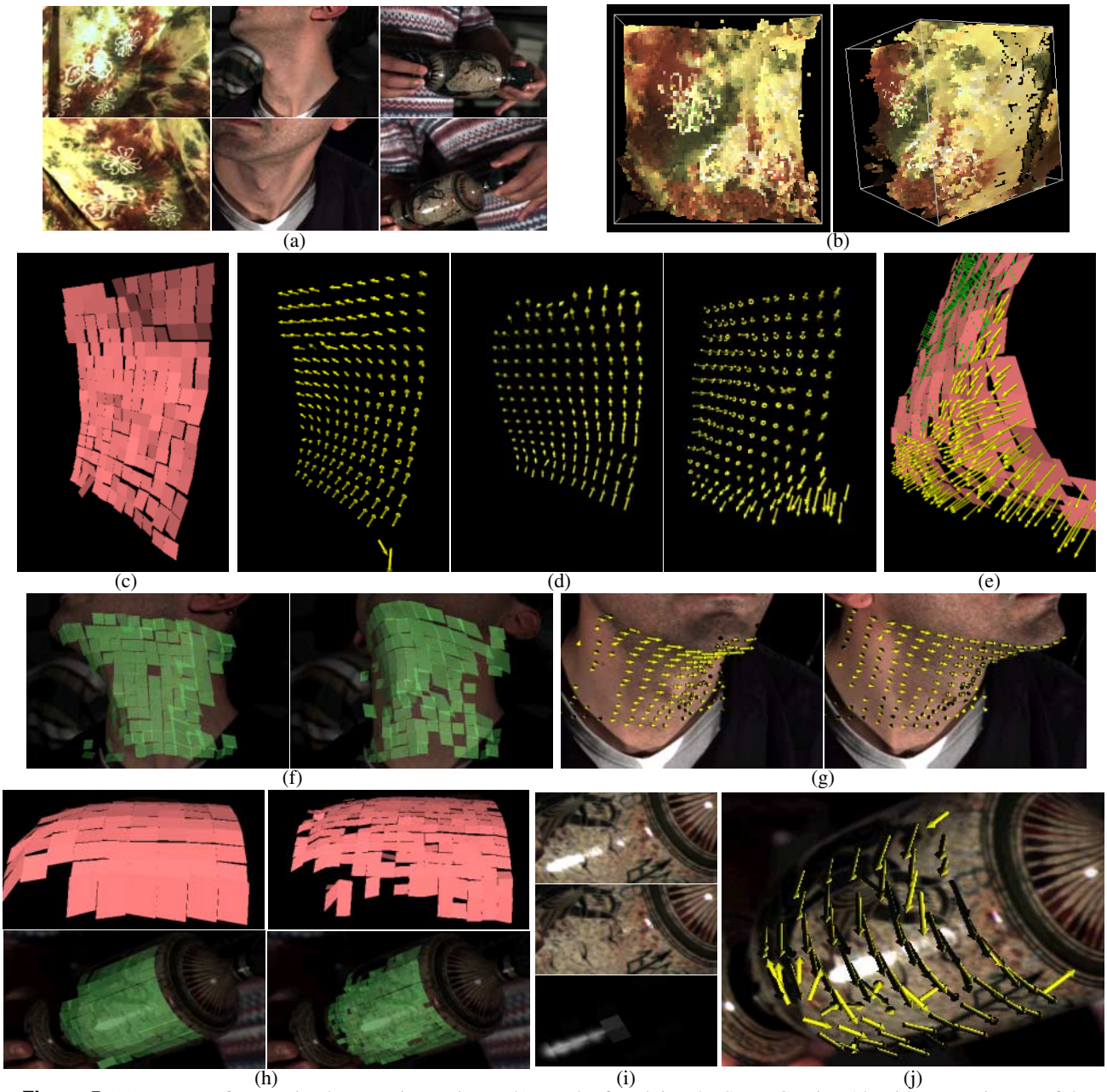

Figure 5. (a) Two out of seven simultaneous input views. (b) Result of applying the Space Carving Algorithm to an instant of the flag sequence. The flag's slowly-varying texture results in an extremely conservative reconstruction that bares little resemblance to the flag's true shape. (c) Results of the Surfel Sampling Algorithm on the same images. (d) Reconstructed motion field for t=0, 4 and 14 with one vector per surfel. (e) Closeup view at t=14 with five vectors per surfel. (f) Neck reconstruction overlaid with input views (*left:* camera 2, t=0; *right:* camera 2, t=27). (g) Neck motion field (*left:* camera 3, t=27; *right:* camera 3, t=52). Note that a portion of the overlaid field is actually occluded from camera 3; it is correctly recovered because occlusions are resolved and taken into account during motion processing. (h) Two views of a vase reconstruction for two resolutions of the cell array (*left:* 16x12x8; *right:* 32x24x8). Note that the strong specular highlight, evident in the semi-transparent image+surfel overlays, does not adversely affect the reconstructed shape or its motion field, shown in (j). (i) *Top to bottom:* Close-up view of an input vase image and its decomposition into diffuse and specular components, respectively. (j) Reconstructed instantaneous motion field. See *http://www.cs.rochester.edu/u/kyros/site/research/motion/surfels.html* for more results and videos.

[27] R. L. Carceroni and K. N. Kutulakos, "Multi-view scene capture by surfel sampling," URCS TR #747, University of Rochester, 2001.

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

[29] R. L. Carceroni and K. N. Kutulakos, "Multi-view 3D shape and motion recovery on the spatio-temporal curve manifold," in *Proc. 7th ICCV*, pp. 520–527, 1999.

[30] J.-Y. Bouguet and P. Perona, "3D photography on your desk," in *Proc. 6th ICCV*, pp. 43–50, 1998.