Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction

Paulo F. U. Gotardo[1], Tomas Simon[2], Yaser Sheikh[2], and Iain Matthews[1,2]

[1]Disney Research
{paulo.gotardo, iainm}@disneyresearch.com

[2]Carnegie Mellon University
{tsimon, yaser}@cs.cmu.edu

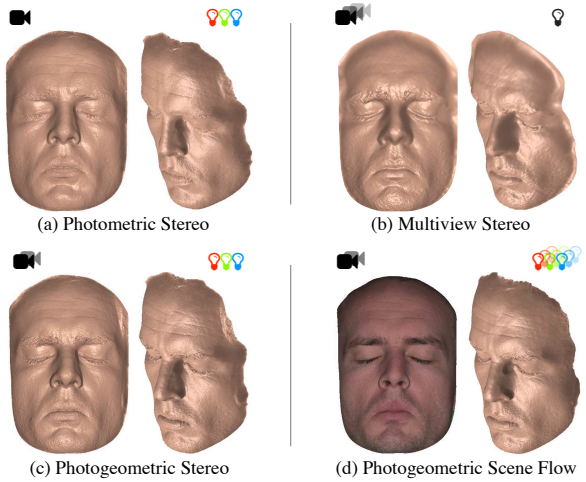
Abstract

Photometric stereo (PS) is an established technique for high-detail reconstruction of 3D geometry and appearance. To correct for surface integration errors, PS is often combined with multiview stereo (MVS). With dynamic objects, PS reconstruction also faces the problem of computing optical flow (OF) for image alignment under rapid changes in illumination. Current PS methods typically compute optical flow and MVS as independent stages, each one with its own limitations and errors introduced by early regularization. In contrast, scene flow methods estimate geometry and motion, but lack the fine detail from PS. This paper proposes photogeometric scene flow (PGSF) for high-quality dynamic 3D reconstruction. PGSF performs PS, OF, and MVS simultaneously. It is based on two key observations: (i) while image alignment improves PS, PS allows for surfaces to be relit to improve alignment; (ii) PS provides surface gradients that render the smoothness term in MVS unnecessary, leading to truly data-driven, continuous depth estimates. This synergy is demonstrated in the quality of the resulting RGB appearance, 3D geometry, and 3D motion.

1. Introduction

High-resolution geometry and appearance are invaluable assets in the movie and video game industries—the quality of a 3D model can make or break the perceived realism of an animation. When it comes to facial models, people have a remarkably low threshold for inaccuracies, and even the smallest details in geometry and appearance are important. In this paper, we improve the resolution and detail in the geometry and appearance of dynamic 3D capture.

Photometric stereo (PS) is a well-established technique to capture high-detail geometry and appearance of real objects, observed under different illumination conditions [32], see Fig. 1(a). PS is generally used to enhance the detail of an initial low-resolution geometry, most often obtained via multiview stereo (MVS) [11,21]. Although computed independently, PS and MVS are complementary in nature: PS provides continuous depth values (fine detail) even for textureless 3D surfaces, but suffers from integration drift (relative depth); MVS provides absolute depth, but its estimates suffer from matching errors and spatial regularization further smoothes fine detail, Fig. 1(b). When PS and MVS are computed independently, this synergy is under-explored [8].



(a) Photometric Stereo (b) Multiview Stereo (c) Photogeometric Stereo (d) Photogeometric Scene Flow

Figure 1. Photogeometric scene flow (PGSF) provides high-quality 3D surface, motion and RGB albedo by simultaneously solving photometric stereo (PS), multiview stereo (MVS) and optical flow (OF). The usual combination of color-PS and MVS (c) corrects for low-frequency shape errors, but presents underconstrained shadow regions (*e.g.*, nose) and monochromatic albedo.

In dynamic PS reconstruction, object motion introduces the additional need to align images taken under significant changes in illumination, which remains a challenging problem. Typically, optical flow (OF) [7,16] is used as a third independent component, with residual misalignment leading to loss of detail [31]. As an alternative, 3-color PS [13,33] was proposed to instantaneously capture 3D geometry in a single frame—with spectrally multiplexed illumination. However, 3-color PS requires objects to be monochromatic and also suffers from self-shadowing (missing data) [14]. Fig. 1(c) is an example color-PS+MVS reconstruction.

In this paper, we propose a new approach for dynamic 3D capture that simultaneously computes PS, MVS, and OF in a coupled way. Key to this approach is the fact that PS not

only *benefits* from, but also *facilitates* the computation of MVS and OF. Together, the solution of these subproblems provides highly detailed 3D geometry, appearance, and instantaneous 3D motion (flow). We therefore refer to this approach as photogeometric scene flow (PGSF). As in scene flow [26], PGSF provides dense estimates of 3D *motion* but at an increased level of *geometric detail* and with relightable, full-color object *appearance*, see Fig. 1(d).

The following sections introduce PGSF using a simple acquisition setup with two cameras and 9 directional lights of 3 different colors, Fig. 2. These lights are multiplexed in time and spectrally to provide an adequate sampling of the instantaneous appearance of a deforming object, within a short interval of only 3 video frames. This minimizes the need for motion compensation, while avoiding overly restricting the reflectance model due to insufficient sampling (the main problem in 3-color PS). We then present a PGSF algorithm for dynamic 3D reconstruction and motion compensation under rapid, controlled changes in illumination. Our new approach is general and can be used with large-scale lighting setups (*e.g.* [11, 20]) and can adopt the latest advances in optical flow theory and stereo matching.

2. Related Work

As the literature on PS, MVS, OF and scene flow is extensive, we focus on previous work closely related to PGSF.

A common approach to video-based, dynamic 3D capture is MVS [10, 36]. Because disparity (depth) is ambiguous in regions with little salient texture, MVS results are regularized and do not capture the finer surface details. Thus, state-of-the-art approaches [6, 25] introduce photometric constraints to partially recover missing detail and improve visual perception of results. However, variability in illumination is insufficient to define surface gradients (normal vectors) unambiguously. The recovered appearance also includes undesirable baked-in shading effects. Similarly, scene flow algorithms [26] are based on MVS [4] and RGB-D sensors [23] that provide reduced geometry detail.

In standard, time-multiplexed PS, a single camera can be used to obtain multiple images of a static object under varying illumination [32]. These images are used to recover the relative depth of pixels revealing the fine detail of 3D surfaces, even for uniform or texture-free surfaces that are impossible to correspond in multiple views. As absolute depth is unconstrained, reconstructions can present arbitrary low-frequency (surface depth integration error) deformations. Thus, PS is often used to enhance the detail of low-resolution geometry obtained independently with MVS [11, 27]. The benefits of simultaneous computation of MVS and PS are investigated in [8], but only for static objects. To make the problem convex, their method approximates depth as a piecewise linear function.

When dynamic objects, such as faces, are captured with time-multiplexed PS, image alignment is required before reconstruction. As the brightness constancy assumption of most image alignment methods does not hold in PS, this task becomes quite challenging. Existing work on image alignment under varying illumination often assume gradual, global linear change [2], small relative motion of a light source [5, 35], or other shading variations [29] that are relatively small when compared to illumination changes in PS. For face capture, [30] interleaves special, uniformly-lit tracking frames in the acquisition process; OF is computed between pairs of tracking frames and interpolated for the frames in between, assuming linear motion. Artifacts resulting from this linear assumption are demonstrated by [31]. They propose the use of complementary illumination conditions that can be simultaneously aligned with a tracking frame without interpolation. Their method is limited to large-scale gradient illumination and can suffer from self-shadowing in smaller setups. The need for special tracking frames also limits temporal resolution.

Color PS with 3 colored lights does not require motion compensation because it can capture three different illumination conditions simultaneously on an RGB sensor [13, 14, 33]. However, 3-color PS cannot recover both surface gradients and RGB albedo (only three inputs for five unknowns). Also, unavoidable self-shadow areas remain underconstrained, require regularization, and introduce reconstruction artifacts due to concave-convex ambiguities [15]. Most variants of 3-color PS only address the former problem, imposing a monochromaticity constraint on the observed object. An undesired consequence is that light calibration becomes object-dependent, more difficult to compute [28], and incorrect when chromaticity varies. In [1], the chromaticity assumption is relaxed to piecewise constant, but the method may introduce segmentation errors on objects with complex appearance variations over space and time. A simple solution used in [18] is to cover a face in white make-up, at the cost of losing fine surface detail and the ability to recover face albedo.

Other color PS approaches avoid the monochromaticity assumption by acquiring more than one image (3 color channels). In [12], a 6-channel image is obtained in a single shot using two cameras aligned with a beam splitter, color filters, and six spectrally distinct light sources. The spectral distribution of surface reflectance is assumed to lie within a low-dimensional space, requiring scene-dependent calibration to alleviate reconstruction bias. In [17], color- and time-multiplexed illumination are combined to reconstruct dynamic surfaces. Image alignment is done via OF in the red channel, with constant frontal illumination. Their approach uses five colored lights and yields surface estimates only at every other video frame. In both [12] and [17], as in 3-color PS, the object is illuminated from only three directions and the shadow problem remains unaddressed.

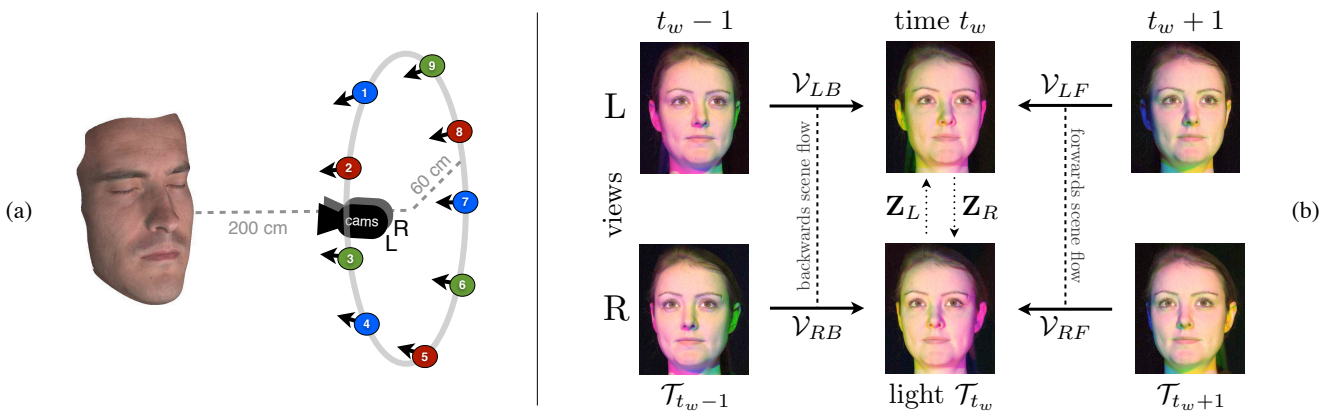


Figure 2. Binocular PGSF with time- and color-multiplexed illumination: (a) setup of camera and colored light sources used to acquire 9 illumination conditions in every 3-frame video interval; (b) the PGSF reconstruction problem is formulated as recovering the unknown surface at time t_w and its backwards and forwards motion within the 3-frame sliding window. These unknowns are parameterized by two depthmaps and four 2D flow fields. Images are aligned by warping in the opposite direction of motion.

In contrast to the previous work, we propose a simple setup that combines color and time multiplexing to obtain a richer sampling on 9 different directions of illumination. This setup reduces not only shadowing, but also the amount of regularization and motion compensation. Photometric calibration is simple and does not require any additional assumption on diffuse reflectance. PGSF addresses the problem of image alignment under rapid changes in illumination without requiring assumptions such as linear motion. The result is full RGB albedo, 3D geometry and motion without penalizing spatial detail or temporal resolution.

3. Photogeometric Scene Flow

This section presents an overview of photogeometric scene flow (PGSF) for dynamic, detailed 3D reconstruction from video. As discussed above, current state-of-the-art approaches correspond to combinations of PS, MVS and OF. Most often, these difficult subproblems are solved independently, with results combined in a final stage that accumulates intermediate errors. Here, we propose PGSF as the simultaneous and synergistic solution of PS, MVS, and OF to overcome challenges faced when these problems are solved independently. While it is already clear that PS benefits from OF alignment and from absolute depth from MVS, these relations are in fact mutually beneficial because:

1. PS equips MVS with knowledge of surface gradients, allowing for truly *data-driven stereo matching* with continuous disparities and no need for spatial regularization, which could over-smooth detail (Sec. 4).

2. PS facilitates OF under rapid, significant changes in illumination; knowledge of illumination and surface geometry (normal vectors) can be used to *relight input images* to closely resemble each other (Sec. 5).

These relations are dependent on an adequate sampling of surface reflectance to unambiguously determine surface orientation and RGB appearance in each video frame. This requirement guides our choice of reflectance model and acquisition setup described next.

3.1. PGSF With 9 Colored Lights

The capture of dynamic events, such 3D facial expressions, is restricted by the short time window to sample instantaneous surface geometry and appearance. For this reason, we adopt a simple Lambertian model with five degrees of freedom (normal and RGB albedo). Thus, highlights (and shadows) are outliers that must be detected and ignored [3]. Highlights can also be filtered out by cross-polarizing light and sensor [20].

To fit the Lambertian model each surface patch (pixel) must be observed under, at least, five different illumination directions. Using color- and time-multiplexed illumination, at least two consecutive video frames (6 color channels) are required. This fact introduces the need for adequate motion compensation. Due to self-shadows, just two frames are often insufficient to provide enough samples for many surface patches of non-convex shapes – *i.e.*, regions around the nose and cheeks on a human face. Furthermore, using a minimal number of observations makes the results more sensitive to sensor noise and other imaging artifacts. For these reasons, we propose a setup with 9 colored light sources to sample reflectance under a richer set of directional illuminations, within a 3-frame time window. This setup is illustrated in Fig. 2(a). Note that there is no increase in the complexity of 3-frame PGSF reconstruction, *versus* the 2-frame case, since motion compensation is only required between adjacent frames (now for both directions in time). The result is more robust 3D reconstruction.

For our experiments, we built a light rig with 9 small clusters of LEDs of a single color—referred to using labels *red*, *green*, or *blue*—approximately matching the peak wavelength sensitivity of each of the RGB camera sensors. These sources are mounted in a nearly circular configuration, Fig. 2(a). We define triplets of RGB sources $\mathcal{T}_t$ (*e.g.*, $\mathcal{T}_t \in \{[5,6,1],[2,9,4],[8,3,7]\}$) that are turned on one at a time in a sequential manner, each for the duration of one video frame (please see supplementary video). These RGB triplets can be defined in different ways. However, it is important to spread out light sources (in space and time) to minimize the intersection of their shadow areas on the target object. This substantially reduces missing data (shadows).

As MVS is required to avoid low-frequency deformations in PS reconstruction, we consider the simplest case with a binocular setup in which the two synchronized cameras observe the target object under the same illumination condition, which varies over time. Setups with more than two cameras are possible. Video acquisition is performed in a dark studio, with negligible ambient light. An example 3-frame window is shown in Fig. 2(b).

The goal of PGSF is to recover the 3D geometry and RGB albedo of the surface at each time t. This task is performed sequentially with a 3-frame sliding window centered at the current frame t_w in both camera views. To make full use of the photometric constraints in this time window, image alignment is required. Thus, we need to estimate (and remove) the surface motion between frames (t_w, t_w+1) and (t_w, t_w-1). Due to nonrigid motion, possible (dis)occlusion and the discrete nature of images, we differentiate the backwards motion (t_w, t_w-1) from the forwards motion (t_w-1, t_w). Alignment warps images in the direction opposite to motion, Fig. 2(b).

The unknowns are encoded simply and effectively on the image grids of each view, left (L) and right (R), at time t_w. The 3D surface is parameterized by depthmaps $\mathbf{Z}_L$ and $\mathbf{Z}_R$. Forwards and backwards 3D motions are each encoded by two 2D optical flow fields (one per view) with cross-view consistency constraints to account for the extra degree of freedom. The forwards/backwards vector fields are denoted $\mathcal{V}_{LF}$, $\mathcal{V}_{RF}$, and $\mathcal{V}_{LB}$, $\mathcal{V}_{RB}$.

Formally, the PGSF objective is defined in terms of the unknowns $\mathcal{X}_c = \{\mathbf{Z}_c, \mathcal{V}_{cB}, \mathcal{V}_{cF}\}$, for $c \in \{L, R\}$, as

$$\min_{\mathcal{X}_L, \mathcal{X}_R} E_{surf}(\mathcal{X}_L, \mathcal{X}_R) + E_{flow}(\mathcal{X}_L, \mathcal{X}_R). \qquad (1)$$

These surface and flow energies represent, respectively, the mutually beneficial relations between PS-MVS (Sec. 4) and PS-OF (Sec. 5). They are minimized in a coupled manner, and in alternation, following the common coarse-to-fine approach with Gaussian image pyramids (Sec. 6).

Estimating the parameters above requires establishing pixel correspondences across views and time. With rapid and significant changes in illumination, the traditional assumption of constant pixel brightness needs to be revised. A more adequate assumption is that RGB albedo remains locally consistent over time (*i.e.*, adjacent video frames). To derive the surface and flow energies, we first formally define our Lambertian model and basic albedo constraint.

3.2. Albedo Consistency

Consider a Lambertian surface patch p with RGB albedo $\boldsymbol{\alpha}_p = [\alpha_p^r, \alpha_p^g, \alpha_p^b]$ and normal vector $\mathbf{n}_p$ at a particular time t. This patch is simultaneously illuminated by three directional lights of distinct colors ($\mathcal{T}_t$). Let $\mathbf{l}_r$, $\mathbf{l}_g$, and $\mathbf{l}_b \in \mathbb{R}^3$ denote their light direction vectors scaled by the corresponding light intensity. For simplicity of notation, we represent normal and light vectors within a camera's local coordinate system; time and camera indices are omitted when defining general relations.

After acquisition and demultiplexing of color channels, this patch is depicted by an RGB pixel $\mathbf{i}_p = [i_p^r, i_p^g, i_p^b]^T$,

$$\mathbf{i}_p = \mathbf{M}^{-1}\hat{\mathbf{i}}_p = \begin{bmatrix} \alpha_p^r & & \\ & \alpha_p^g & \\ & & \alpha_p^b \end{bmatrix} \begin{bmatrix} \mathbf{l}_r^T \\ \mathbf{l}_g^T \\ \mathbf{l}_b^T \end{bmatrix} \mathbf{n}_p, \qquad (2)$$

where $\hat{\mathbf{i}}_p$ has the captured, multiplexed color values. Each column of the mixing matrix $\mathbf{M} = [\mathbf{m}_r, \mathbf{m}_g, \mathbf{m}_b] \in \mathbb{R}^{3\times3}$ encodes the RGB color of a light source as seen by the camera (including color space transformations applied by camera firmware). Here, the illumination parameters $\mathbf{M}$, $\mathbf{l}_r$, $\mathbf{l}_g$, and $\mathbf{l}_b$ are pre-calibrated for each triplet of LEDs, $\mathcal{T}_t$, using a white diffuse sphere and a mirrored sphere as in [14]. The unknowns in (2) are therefore $\boldsymbol{\alpha}_p$ and $\mathbf{n}_p$.

To derive the albedo consistency constraint, consider a pair of corresponding pixels (p, p') from two images $(1, 2)$ across time or cameras; their values are denoted $i_{p,1}^\lambda$ and $i_{p',2}^\lambda$, for each color channel $\lambda \in \{r, g, b\}$. From (2), we define the basic pairwise image relation in PGSF as:

$$\alpha_{p,1}^\lambda \approx \alpha_{p',2}^\lambda \quad \Rightarrow \quad \frac{i_{p,1}^\lambda}{\mathbf{l}_{\lambda,1}^T \mathbf{n}_p} \approx \frac{i_{p',2}^\lambda}{\mathbf{l}_{\lambda,2}^T \mathbf{n}_p}. \qquad (3)$$

To simplify notation, λ is omitted in the following sections, but we always consider pairs of the same color channel. When the two images are taken from adjacent time instants (t, t'), this relation also implies the standard assumption in time-multiplexed PS, $\mathbf{n}_{p,t} \approx \mathbf{n}_{p',t'}$. This assumption corresponds to an as-rigid-as-possible, small motion model that is common in optical/scene flow methods [26].

The relation in (3) is completely defined by motion (correspondence) and geometry ($\mathbf{n}_p$), without actually requiring explicit albedo estimates. This fact eliminates a large number of unknowns, since per-pixel RGB albedo does not need to be computed before the PGSF solution has been obtained.

4. Surface Reconstruction

The surface estimation step performs simultaneous PS and binocular stereo, given current estimates of the 2D flows (pre-aligned images). Thus, surface estimation at time t can be formulated as a rigid problem once motion has been removed from the input.

Next, we derive the unary and binary surface energy terms on $\mathcal{X}_c$. To emphasize that flows are treated as constants during surface update, we express these energies only in terms of depthmap $\mathbf{Z}_c$ in each camera view,

$$E_{surf}(\mathcal{X}_L, \mathcal{X}_R) = \sum_c E_{PS}(\mathbf{Z}_c) + \beta_1 E_{SM}(\mathbf{Z}_c) + \beta_2 E_{LR}(\mathbf{Z}_L, \mathbf{Z}_R). \quad (4)$$

These are the PS, stereo matching (SM), and LR-consistency energies, weighted by constants β_1 and β_2.

To take full advantage of the complementary nature of PS and MVS, we express all constraints directly in terms of $\mathbf{Z}_c$, avoiding the need to compute intermediary results (normal vectors and disparity) that can propagate errors.

4.1. Photometric Stereo

Consider the basic PGSF relation in (3) and a pair of images $t, t' \in \{t_w - 1, t_w, t_w + 1\}$ in the same view c. Cross-multiplying denominators, we rewrite (3) as

$$\underbrace{\left(i_{p't'}\mathbf{l}_t^T - i_{pt}\mathbf{l}_{t'}^T\right)}_{[a_{ptt'}\ b_{ptt'}\ c_{ptt'}]}\mathbf{n}_p \approx 0, \quad (5)$$

where the constant vector $[a_{ptt'}\ b_{ptt'}\ c_{ptt'}]$ is orthogonal to $\mathbf{n}_p$. We now relate this linear constraint to the gradient field ∇z_p of depth values $z_p = \mathbf{Z}_c(u_p, v_p)$, where (u_p, v_p) are the coordinates of pixel p on the image grid of view c.

In perspective PS, we have

$$\mathbf{n}_p \propto \begin{bmatrix} 1 & & \\ & 1 & \\ \frac{u_p}{f_c} & \frac{v_p}{f_c} & \frac{1}{f_c} \end{bmatrix} \begin{bmatrix} \nabla z_p \\ z_p \end{bmatrix}, \quad (6)$$

where f_c is the focal distance of camera c and $\propto$ denotes up-to-scale equality. Equation (6) is an equivalent expression of a proof given in [22]. Since (5) is up-to-scale, it can be combined with (6) to yield the photometric energy term

$$E_{PS}(\mathbf{Z}_c) = \sum_{p,t,t',\lambda} \left(\begin{bmatrix} a_{ptt'} \\ b_{ptt'} \\ c_{ptt'} \end{bmatrix}^T \begin{bmatrix} 1 & & \\ & 1 & \\ \frac{u_p}{f_c} & \frac{v_p}{f_c} & \frac{1}{f_c} \end{bmatrix} \begin{bmatrix} \nabla z_p \\ z_p \end{bmatrix} \right)^2, \quad (7)$$

over all pixels p and pairs of images (t, t'). Because ∇z_p is a linear function of $\mathbf{Z}_c$ (using forwards derivatives), minimizing (7) is equivalent to solving a large and sparse linear system of (homogeneous) equations on the depthmap $\mathbf{Z}_c$.

Considering all pairs of images in our 3-frame window, and all three RGB color channels (λ), we obtain 6 independent constraints on the 2 degrees of freedom of $\mathbf{n}_p$ (or ∇z_p). This overconstraining is welcome since, in practice, the effective number of constraints may be reduced by self-shadowing in one or more color channels. It is still possible for a very small number of pixels to present only 1 or 0 pairwise constraints (*e.g.*, points near the nostril cavities on a face). Only in these rare cases we adopt the curvature-based anisotropic regularization method of [15].

4.2. Stereo Matching

The basic PGSF constraint is now applied to image pairs (c, c') at the same time t but across views (to simplify notation, we also omit the time index t). Both images present illumination $\mathcal{T}_t$, so we can drop the denominators in (3). Let $\mathbf{u}_p = (u_p, v_p)$ denote pixel coordinates of p. At each time instant in the 3-frame window, we have

$$i_c(\mathbf{u}_p) = i_{c'}(\mathbf{u}_{p'}), \quad \mathbf{u}_{p'} = \mathbf{u}_p + \mathbf{w}_p, \quad (8)$$

where $\mathbf{w}_p \in \mathbb{R}^2$ is a displacement along the epipolar line. By expressing $\mathbf{w}_p = w(z_p)$ in terms of $z_p = \mathbf{Z}_c(\mathbf{u}_p)$ and camera calibration, we can perform stereo matching while solving directly for the depthmap. The result is a set of non-linear constraints on $\mathbf{Z}_c$ that need to be enforced iteratively. Nevertheless, this task can be incorporated naturally into the coarse-to-fine optimization strategy of PGSF (Sec. 6).

The advantages of this approach are two-fold: (i) we directly triangulate continuous depth values; (ii) the new stereo matching constraints on z_p (absolute depth) and the previous photometric constraints on ∇z_p (relative depth, detail) complement each other naturally; they render it unnecessary to define a spatial smoothness term for stereo matching (which could blur surface detail).

The 2D displacement vector along to the epipolar line is

$$w(z_p) = \frac{\mathbf{A}[\mathbf{u}_p^T\ 1]^T z_p + \mathbf{b}}{\mathbf{c}^T[\mathbf{u}_p^T\ 1]^T z_p + d} - \mathbf{u}_p, \quad \mathcal{M}_{c'c} = \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{c}^T & d \end{bmatrix}, \quad (9)$$

$\mathcal{M}_{c'c} \in \mathbb{R}^{3\times 4}$ is the projection matrix from view c to c' (including camera intrinsincs), and $\mathbf{A} \in \mathbb{R}^{2\times 3}$.

The stereo matching energy term is then defined as

$$E_{SM}(\mathbf{Z_c}) = \sum_{p,t,\lambda} \left(i_{c'}(\mathbf{u}_p + w(z_p)) - i_c(\mathbf{u}_p)\right)^2 \gamma_p, \quad (10)$$

where γ_p is an occlusion weight based on foreshortening and consistency of pairs (p, p') given by $\mathbf{Z}_L$ and $\mathbf{Z}_R$ [34].

During optimization, the energy $E_{SM}(\mathbf{Z_c})$ is linearized about the current estimate of each z_p. This yields a linear data term (gradient) for the surface update Δz_p,

$$\nabla i_{c'}(\mathbf{u}_{p'})^T \mathbf{J}_w \Delta z_p = i_c(\mathbf{u}_p) - i_{c'}(\mathbf{u}_{p'}), \quad (11)$$

where $\mathbf{J}_w \in \mathbb{R}^2$ is the Jacobian of $w(z_p)$ in (9).

4.3. Cross-View Consistency

To enforce that $\mathbf{Z}_L$ and $\mathbf{Z}_R$ provide a consistent representation for those surface patches that are visible on both views, we define the additional energy term,

$$E_{LR} = \sum_{c \neq c'} \sum_{p} (\mathbf{c}^T [\mathbf{u}_p^T \; 1]^T z_p^c + d - z_p^{c'})^2 \gamma_p, \qquad (12)$$

based on the projection matrices $\mathcal{M}_{c'c} \in \mathbb{R}^{3\times 4}$ in (9). By optimizing $\mathbf{Z}_L$ and $\mathbf{Z}_R$ in an alternated manner, the energy in (12) can also be treated as a set of linear constraints.

5. Motion Estimation

Motion estimation is used to achieve image alignment, over time, in each camera view. This step is based on a novel algorithm for the challenging task of performing OF under significant illumination change. The key idea is to take advantage of known surface normals, and illumination $\mathcal{T}_t$, to *relight input images* to closely resemble each other.

The OF algorithm derived next follows the seminal, variational approach of [16], with spatial regularization to address the aperture problem. We assume *albedo* consistency over time, rather than brightness. The general method that we propose can be adapted easily to take advantage of new developments in optical/scene flow estimation.

Given the current depthmap estimates, we define the unary and binary motion energies in terms of the four 2D flows $\mathcal{V}_{c\delta}$, for $c \in \{L, R\}$ and direction $\delta \in \{B, F\}$,

$$E_{flow}(\mathcal{X}_L, \mathcal{X}_R) = \sum_{c,\delta} E_{PF}(\mathcal{V}_{c\delta}) + \beta_3 E_{TV}(\mathcal{V}_{c\delta}) \qquad (13)$$
$$+ \sum_{\delta} \beta_4 E_{SF}(\mathcal{V}_{L\delta}, \mathcal{V}_{R\delta}),$$

where β_3 and β_4 are fixed energy weights. The energy $E_{TV}(\cdot)$ represents the typical total variation (TV-L1) regularizer [7]. The photometric flow (PF) and scene flow (SF) energies are derived as follows.

5.1. Photometric Flow

Consider the basic relation (3) for pairs of images in the same view (index c is omitted) but at different times, with $t = t_w$ and $t' \in \{t_w - 1, t_w + 1\}$. With known illumination and normal vectors, we can pre-compute the shading terms in the denominators as the scalar $s_{pt} = \mathbf{l}_t^T \mathbf{n}_p$. From (3), we obtain the relighting relation,

$$s_{pt'} i_t(\mathbf{u}_p) \approx s_{pt} i_{t'}(\mathbf{u}_{p'}), \quad \mathbf{u}_{p'} = \mathbf{u}_p + \mathbf{w}_p, \qquad (14)$$

defined between two *cross-shaded images*. The coordinates of $\mathbf{u}_{p'}$ are defined by the 2D flow $\mathbf{w}_p \in \mathcal{V}_{c\delta}$. Note that the surface shading terms are already defined on the image grid at time t. Thus, only the image at time t' is warped.

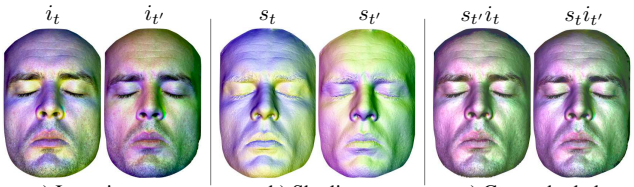


Figure 3. Relighting for photometric flow: images (a) are cross-shaded (c) to closely match each other during alignment. Shading values (b) are given by the surface estimate and light calibration.

This relighting relation can be understood as using the known illumination patterns $\mathcal{T}_t$ and $\mathcal{T}_{t'}$ to relight i_t and i'_t to closely match each other. Figure 3 illustrates this operation. Clearly, relighting facilitates OF computation since differences in illumination are canceled. Another desirable effect is down-weighting the influence of pixels with small shading values (*i.e.*, high photometric foreshortening).

Note that shadowed pixels represent missing input and, thus, cannot be relit. Fortunately, a pixel is rarely shadowed in more than one color channel with spectrally multiplexed illumination. The flow $\mathbf{w}_p$ is still constrained by the other channels and, to a lesser extent, by the TV-L1 regularizer.

The relighting relation incorporates an image warp (the 2D flow field $\mathbf{w}_p$) and reflects a fundamental fact: image alignment improves photometric reconstruction, and *vice versa*. For $\mathbf{w}_p \in \mathcal{V}_{c\delta}$, the photometric flow energy is the alignment error between cross-shaded images,

$$E_{PF}(\mathcal{V}_{c\delta}) = \sum_{p,\lambda} (s_{pt} i_{t'}(\mathbf{u}_p + \mathbf{w}_p) - s_{pt'} i_t(\mathbf{u}_p))^2. \qquad (15)$$

As in standard optical flow, the image indexing operation results in nonlinear energy minimization. Linearizing (15) yields one constraint (per pixel) on the update $\Delta\mathbf{w}_p \in \mathbb{R}^2$,

$$s_{pt} \nabla i_{t'}(\mathbf{u}_{p'}) \Delta\mathbf{w}_p + (s_{pt} i_{t'}(\mathbf{u}_{p'}) - s_{pt'} i_t(\mathbf{u}_p)) \approx 0. \qquad (16)$$

This result is a new, relit form of the space-time gradient in standard OF. In fact, one can demonstrate that it represents a weighted version of the constraint $\frac{\mathrm{d}\alpha_p}{\mathrm{d}t} \approx 0$.

5.2. Scene Flow

The scene flow energy term encourages the 2D flows $\mathcal{V}_{L\delta}$ and $\mathcal{V}_{R\delta}$ to be consistent with a 3D vector field, thus fixing the extra degree of freedom (per pixel) in the parameterization, which could introduce errors into the estimated flows.

Let (p, p') denote corresponding pixels across views, defined by the depthmaps $\mathbf{Z_c}$. Their 2D flow vectors are $\mathbf{w}_p \in \mathcal{V}_{L\delta}$ and $\mathbf{w}_{p'} \in \mathcal{V}_{R\delta}$. These pixels and displacements are related by the projection $\mathcal{M}_{c'c}$ in (9), yielding

$$\mathbf{w}_{p'} = \frac{\mathbf{A}[(\mathbf{u}_p + \mathbf{w}_p)^T \; 1]^T (z_p + w_p^z) + \mathbf{b}}{\mathbf{c}^T [(\mathbf{u}_p + \mathbf{w}_p)^T \; 1]^T (z_p + w_p^z) + d} - \mathbf{u}_{p'}. \qquad (17)$$

Algorithm 1 Photogeometric Scene Flow (PGSF)

for 3-frame window at time t **do**
  Detect shadows, highlights on the 6 input images.
  Compute a Gaussian pyramid for each image.
  Initialize: $\mathcal{V}_{\{L,R\}\{B,F\}} \leftarrow$ nil,
  $\mathbf{Z}_{\{L,R\}} \leftarrow$ orthographic PS at coarsest resolution (pyramid top).
  for each pyramid level, from coarse to fine, **do**
    Update flows $\mathcal{V}_{\{L,R\}B}$ in alternation, given $\mathbf{Z}_{\{L,R\}}$ (Sec. 5).
    Update flows $\mathcal{V}_{\{L,R\}F}$ in alternation
    Align images $\mathbf{i}_{t-1}, \mathbf{i}_t, \mathbf{i}_{t+1}$ in each view.
    Update depthmaps $\mathbf{Z}_{\{L,R\}}$ in alternation (Sec. 4).
  end for
  Compute final RGB albedo at time t using Eq. (2).
end for

Equation (17) provides two linear constraints on the unknown scalar w_p^z, the third component of the 3D scene flow represented in view c. Therefore, after each update of the 2D flows, we can compute w_p^z and enforce that the 2D flows are consistent with the projections of this single 3D scene flow. We enforce this consistency symmetrically, on both views, by minimizing the scene flow energy

$$E_{SF}(\mathcal{V}_{L\delta}, \mathcal{V}_{R\delta}) = \sum_{p,c} \|\mathbf{w}_p - \mathcal{P}(\mathbf{w}_{p'}, w_{p'}^z)\|_2^2 \gamma_p. \quad (18)$$

The projection $\mathcal{P}(\cdot)$ is as in (17); γ_p is the occlusion weight.

6. Optimization

To handle large in-plane motion at high resolutions, both surface and motion are estimated in a coarse-to-fine manner using Gaussian image pyramids [9]. We adopt block-coordinate descent optimization and update surface and motion estimates (Sec. 4 and 5) in alternation. In addition, within these steps, the unknowns on the L- and R-views are also updated in alternation for proper minimization of binary energy terms, as listed in Algorithm 1.

The total number of alignment-reconstruction steps is controlled by the number of levels of the Gaussian pyramids. We define the number of levels using a typical 75% image down-sampling factor. At the coarsest level, flow is initialized as nil and the initial surface is obtained from orthographic PS (initial absolute depth is given by camera calibration). Surface reconstruction uses off-the-shelf solvers for large and sparse systems of linear equations (*e.g.*, Mathworks Matlab solvers). Our flow algorithm is based on [7, 19] and uses successive over-relaxation.

Reconstruction is performed with a 3-frame sliding window. It can optionally reuse the result from the previous window to initialize and constrain the subsequent solution, but this is not required. Initial shadow detection is carried out by simple image thresholding [14]. Highlights are removed with polarizers [20] or detected as outliers (those image values that are substantially brighter than the rest of the 16-bit sensor data). The resulting mask of outliers is used to discard all constraints involving those observations.

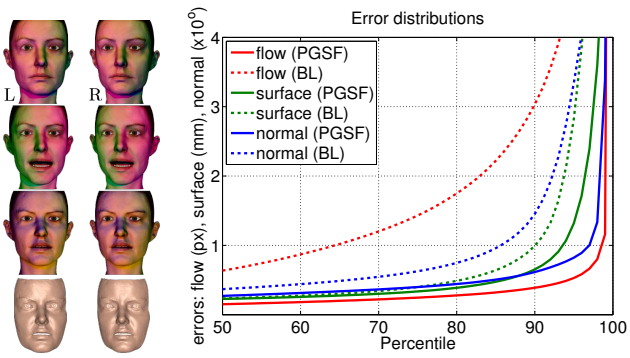


Figure 4. Evaluation on 300-frame synthetic dataset with ground-truth motion and geometry (left) *versus* the baseline method BL.

7. Experimental Results

Our capture setup (Fig. 2) has two cameras located centrally, in front of the light rig, and 1.5 meters from the target object. We use two Sony PMW-F55 cameras (with 85mm lenses) that output 16-bit raw (linear) color at 4k resolution (4096×2160), 60Hz. LED light multiplexing is programmed in an Arduino microcontroller. It is synchronized to camera framerate via an interrupt pin connected to a v-sync decoder and a genlock generator.

For quantitative evaluation, this setup was simulated in software (Maya). A synthetic, 300-frame stereo sequence (5 seconds, 1k resolution) was rendered for a face model animated by real, marker-based motion capture [24]. Dense ground-truth motion and geometry were also generated. We compare the performance of PGSF to that of a baseline algorithm (BL) with a pipeline of three independent steps: (i) OF with linear motion interpolation [30]; (ii) our MVS method with spatial regularization [15]; and (iii) PS with precomputed MVS [21]. Error distributions for estimated motion, surface, and normal vectors are given in Fig. 4. The coupled solutions of PGSF are clearly more accurate, with major improvements concentrated on the 80th-100th error percentiles (due to localized face deformation). For both methods, larger surface errors occur at the sides of the face, where occlusion prevents MVS triangulation. Motion and geometry are also less accurate inside the mouth, due to occlusion and shadowing.

Figure 5 shows the 3D reconstruction for a real video frame acquired with the setup above. Since ground-truth geometry is not available for these real images, we validate surface estimates against the popular PMVS algorithm [10]. PMVS is a state-of-the-art MVS method based on patch matching and does not require regularization, providing 3D point clouds instead of dense depthmaps. On a total of 100 video frames, PGSF and PMVS estimate depth consistently within fractions of a millimeter (Fig. 5, supplementary file). However, PMVS triangulates spurious points at highly foreshortened areas; its results also lack the fine de-

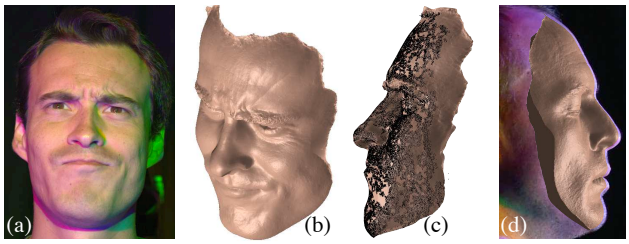


Figure 5. Detailed 3D reconstruction (b) for a real frame (a) in the supplementary video. The profile view (c) is overlaid with the PMVS point cloud, showing close agreement (median depth difference of 0.13 mm, 90^{th} percentile at 0.5 mm, over 100 frames); (d) recovered surface overlaid on image of a profile camera, used only for evaluation (offset added to ease visualization).

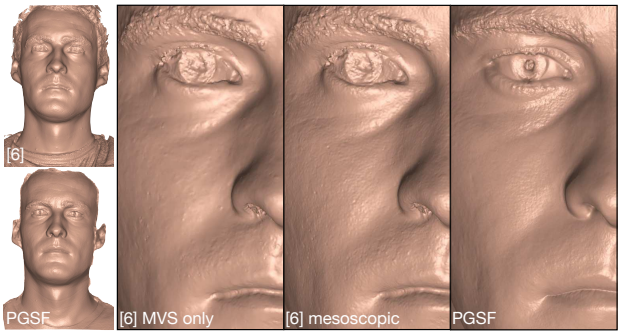


Figure 6. Comparison with the MVS method [6]: results at similar resolution, same actor and expression, different capture sessions.

tail of PGSF. To further assess depth accuracy on real data, Fig. 5(d) shows the reconstructed face overlaid on the actual silhouette seen by a profile camera, used for validation only.

Figure 6 shows the improved detail provided by PGSF for eyes, brows, nostrils and lips in comparison to [6]. With uniform illumination, [6] does not recover shading-free albedo; mesoscopic detail is heuristic and not metrically correct (changes in albedo are mistaken as geometry).

The superior accuracy of cross-shaded photometric flow on real images is also shown in the supplementary video, which compares the residual motion in images aligned by PGSF and by the method in [30]. The supplementary video further demonstrates the quality of the geometry, appearance and 3D motion estimated by PGSF (Fig. 7).

The ability of PGSF to reconstruct highly detailed surfaces with different materials and colors is demonstrated in Fig. 8. An important advantage of PGSF over 3-color PS is the ability to capture temporal variability of RGB albedo, Fig. 9. Full color albedo is a valuable asset in building realistic models for animation and for post-production. For instance, relighting is a frequent task faced by artists during movie and game production in which previously captured performances have to be adapted to match a certain environment. The application of PGSF in the realistic relighting of captured 3D faces is illustrated in the supplementary video.

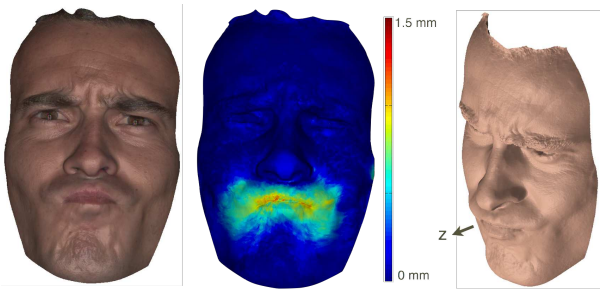


Figure 7. RGB albedo, 3D geometry and estimated motion component orthogonal to the image plane (see supplementary video).

Figure 8. Detailed reconstruction of surfaces with different materials, colors and skin tones: (a)-(b) depthmaps $\mathbf{Z}_L$ and $\mathbf{Z}_R$ (with RGB albedo), (c) recovered surface $\mathbf{Z}_R$ with and without albedo.

Figure 9. Captured temporal variation in RGB albedo due to changes in blood flow (with polarizers, see supplementary video).

8. Conclusion

We propose photogeometric scene flow (PGSF) as the simultaneous and synergistic estimation of PS, MVS, and OF for high-detail, dynamic 3D capture. PGSF couples the solution of these three difficult subproblems to overcome the challenges faced when they are solved independently.

To unambiguously capture surface normals and full RGB albedo in each video frame, we propose a simple binocular setup with 9 colored lights that are spectrally and temporally multiplexed within a period of three frames. This design minimizes shadows and also the amount of motion compensation. Nevertheless, the key ideas in PGSF are general and also applicable in more complex acquisition setups with a different number of cameras and light sources.

References

[1] R. Anderson, B. Stenger, and R. Cipolla. Color photometric stereo for multicolored surfaces. In *Proc. ICCV*, pages 2182–2189, 2011. 2

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Computer Vision*, 56(3):221–255, 2004. 2

[3] S. Barsky and M. Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Trans. PAMI*, 25(10):1239–1252, 2003. 3

[4] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *CVPR*, pages 1506–1513, 2010. 2

[5] R. Basri and D. Frolova. A two-frame theory of motion, lighting and shape. In *Proc. CVPR*, pages 1–7, 2008. 2

[6] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 75:1–75:10, 2011. 2, 8

[7] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*, pages 25–36, 2004. 1, 6, 7

[8] H. Du, D. Goldman, and S. Seitz. Binocular photometric stereo. In *Proc. BMVC*, 2011. 1, 2

[9] D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2003. 7

[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. PAMI*, 32(8):1362–1376, 2010. 2, 7

[11] G. Fyffe, T. Hawkins, C. Watts, W.-C. Ma, and P. Debevec. Comprehensive facial performance capture. In *Eurographics*, 2011. 1, 2

[12] G. Fyffe, X. Yu, and P. Debevec. Single-shot photometric stereo by spectral multiplexing. In *Proc. ICCP*, 2011. 2

[13] C. Hernandez, G. Vogiatzis, G. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *Proc. ICCV*, pages 1–8, 2007. 1, 2

[14] C. Hernandez, G. Vogiatzis, and R. Cipolla. Shadows in three-source photometric stereo. In *Proc. ECCV*, pages 290–303, 2008. 1, 2, 4, 7

[15] C. Hernandez, G. Vogiatzis, and R. Cipolla. Overcoming shadows in 3-source photometric stereo. *IEEE Trans. PAMI*, 33(2):419–426, 2011. 2, 5, 7

[16] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 1, 6

[17] H. Kim, B. Wilburn, and M. Ben-Ezra. Photometric stereo for dynamic surface orientations. In *Proc. ECCV*, pages 59–72, 2010. 2

[18] M. Klaudiny and A. Hilton. High-detail 3D capture and non-sequential alignment of facial performance. In *Proc. 3DIM/3DPVT*, pages 17–24, 2012. 2

[19] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 7

[20] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proc. Eurographics*, pages 183–194, 2007. 2, 3, 7

[21] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Trans. Graph.*, 24(3):536–543, 2005. 1, 7

[22] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *Proc. CVPR*, pages 1474–1481, 2013. 5

[23] J. Quiroga, T. Brox, F. Devernay, and J. Crowley. Dense semi-rigid scene flow estimation from rgbd images. In *ECCV*, 2014. 2

[24] J. R. Tena, F. De la Torre, and I. Matthews. Interactive region-based linear 3D face models. *ACM Trans. Graph.*, 30(4):76:1–76:10, 2011. 7

[25] L. Valgaerts, C. Wu, A. Bruhn, H. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph.*, 31(6):187:1–187:11, 2012. 2

[26] S. Vedula, S. Baker, R. Collins, T. Kanade, and P. Rander. Three-dimensional scene flow. In *Proc. ICCV*, 1999. 2, 4

[27] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 28(5), 2009. 2

[28] G. Vogiatzis and C. Hernandez. Self-calibrated, multi-spectral photometric stereo for 3D face capture. *Int. J. Computer Vision*, 97(1):91–103, 2012. 2

[29] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for TV-L1 optical flow. In D. Cremers, B. Rosenhahn, A. Yuille, and F. Schmidt, editors, *Statistical and Geometrical Approaches to Visual Motion Analysis*, pages 23–45. Springer-Verlag, 2009. 2

[30] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. In *ACM SIGGRAPH 2005 Papers*, pages 756–764, 2005. 2, 7, 8

[31] C. Wilson, A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec. Temporal upsampling of performance geometry using photometric alignment. *ACM Trans. Graph.*, 29(2), 2010. 1, 2

[32] R. Woodham. Photometric method for determining surface orientation from multiple images. *Opt. Eng.*, 19(1):139–144, 1980. 1, 2

[33] R. Woodham. Gradient and curvature from the photometric-stereo method, including local confidence estimation. *J. Opt. Soc. Am. A*, 11(11):3050–3068, 1994. 1, 2

[34] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *IEEE Trans. PAMI*, 31(6):974–988, 2009. 5

[35] L. Zhang, B. Curless, A. Hertzmann, and S. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In *Proc. ICCV*, 2003. 2

[36] L. Zhang, N. Snavely, B. Curless, and S. Seitz. Spacetime faces: High resolution capture for modeling and animation. In *ACM SIGGRAPH 2004 Papers*, pages 548–558, 2004. 2