EUROGRAPHICS 2011 / M. Chen and O. Deussen
(Guest Editors)

Volume 30 (2011), Number 2

Comprehensive Facial Performance Capture

Graham Fyffe[†] Tim Hawkins[‡] Chris Watts[§] Wan-Chun Ma[†] Paul Debevec[†]

[†]Institute for Creative Technologies, University of Southern California
[‡]Lightstage LLC
[§]Bake Visual Effects Inc.

Abstract

We present a system for recording a live dynamic facial performance, capturing highly detailed geometry and spatially varying diffuse and specular reflectance information for each frame of the performance. The result is a reproduction of the performance that can be rendered from novel viewpoints and novel lighting conditions, achieving photorealistic integration into any virtual environment. Dynamic performances are captured directly, without the need for any template geometry or static geometry scans, and processing is completely automatic, requiring no human input or guidance. Our key contributions are a heuristic for estimating facial reflectance information from gradient illumination photographs, and a geometry optimization framework that maximizes a principled likelihood function combining multi-view stereo correspondence and photometric stereo, using multi-resolution belief propagation. The output of our system is a sequence of geometries and reflectance maps, suitable for rendering in off-the-shelf software. We show results from our system rendered under novel viewpoints and lighting conditions, and validate our results by demonstrating a close match to ground truth photographs.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Stereo

1. Introduction

Photorealistic digital faces are becoming increasingly common in entertainment media, due to the compelling storytelling they enable. Producing a believable performance of a fully digital human face, once impossible, is now achievable through significant artistic and technical effort. Major milestones in digital faces include the Universal Capture method used in The Matrix Reloaded [BPL*03], and the facial modeling and animation techniques used in The Curious Case of Benjamin Button, similar to The Digital Emily project [ARL*09]. In this work, we focus on the task of capturing a *comprehensive* digital version of an actor's live facial performance, supporting both novel viewpoints and novel illumination. Applications include face replacement, digital stunt doubles, and integration into virtual sets. We do not consider temporal correspondence or editing the captured performance, which we believe is a research topic in its own right, possibly using a captured performance as input. A plethora of facial animation techniques has been proposed over the years, many with the goal of comprehensive capture [GGW*98, BPL*03, ZSCS04, HWT*04, WGT*05, BLB*08, MJH*08, ARL*09, WGP*10, BHPS10]. The quality of performance capture systems varies dramatically, as does the human effort required to process or use the data after it has been recorded. The highest quality results have resulted from significant human effort, which is out of reach for smaller entities. We propose to raise the bar to consider the following five criteria as *essential* for a comprehensive facial performance capture system: **1. Dynamic capture.** The system must record an actor's performance in real time. **2. Full facial coverage.** The captured geometry must cover the entire face, with some freedom of movement allowed for the actor's head. **3. Detailed geometry.** The captured geometry must be detailed enough to faithfully reproduce occlusion and self shadowing effects when rendered. **4. Detailed reflectance.** The captured reflectance information must be detailed enough to allow photorealistic rendering under any lighting condition, including global illumination effects involving nearby objects. **5. Automatic processing.** After a performance is recorded, the data must be processed automatically, requiring no additional human input or guidance.

We evaluate previous work related to performance capture, noting several approaches that do not meet the five criteria. We then propose a system that aims to meet the five

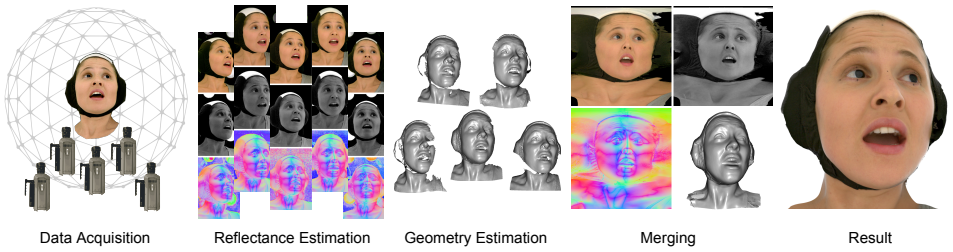

Figure 1: *Overview of our system.*

criteria better than previous work. Figure 1 illustrates an overview of our system, which works as follows: We capture an actor's performance under active gradient illumination using multiple cameras. We then estimate the reflectance function at each pixel in each camera view. Next, we estimate geometry from the point of view of each camera, optimizing with respect to the other views. Finally, we merge the estimates into a single geometry with reflectance maps. This process is repeated for every frame in the performance. We demonstrate our system on a live facial performance, rendered under novel viewpoints and lighting conditions, and conclude with a discussion and directions for future work.

2. Related Work

Computer vision and graphics employ a vast toolbox of 3D measurement techniques which includes passive stereo matching, structured light scanning, photometric stereo algorithms, and shape template methods. Different techniques have different pros and cons, which we evaluate here keeping in mind the five criteria for comprehensive facial performance capture we identified in section 1.

Passive stereo matching. The determination of shape from the stereo correspondence between two images is a well-studied topic; a taxonomy of stereo correspondence approaches is presented in [SS02]. The principal challenges faced by these methods include performance, robust determination of matches near occlusions, detection of matches in the absence of significant texture, and enforcement of smoothness constraints without biasing the recovered correspondences. Some recent stereo techniques which specifically address some of these problems include [SLKS05], which performs symmetric stereo with occlusion handling, [WTRF09], which uses second-order smoothness priors to achieve a more consistent surface, [SZJ09], which uses non-parametric smoothness priors, and most recently [BBB*10], which uses anisotropic second-order smoothness priors to avoid smoothing over depth discontinuities. Many techniques find correspondences between multiple (more than two) views of a scene to increase the quality of the matches. [SCD*06] compares many multi-view stereo algorithms over a collection of benchmark datasets and [GSC*07] shows impressive recent results for community photo collections. Like [GSC*07], [FP09] takes advantage of recent results in automatic structure-from-motion techniques to determine camera positions automatically. Having multiple views improves the likelihood that accurate matches can be found, but it does not dramatically improve the precision of those matches, which remain difficult to resolve at sub-pixel precision. Even the visually pleasing surface details achieved by [BBB*10] are largely cosmetic, as they are hallucinated to match the high spatial frequency statistics of the source photographs, but are not metrically accurate. In terms of comprehensive facial performance capture, passive stereo matching easily meets the dynamic capture, full facial coverage, and automatic processing criteria. It also arguably meets the detailed geometry criterion, thanks to the most recent advances. Indeed, if geometry and a simple color map are the only requirements for an application, then single-shot passive stereo methods are likely sufficient. However, passive stereo matching cannot provide the detailed reflectance information required for comprehensive facial performance capture, nor does it strive to do so. High resolution textures may be captured (as in [BHPS10]), but dependencies on view and lighting direction are completely ignored. On the other hand, passive stereo matching may coexist with many reflectance capture methods, so it is a good candidate for inclusion as a component in our system.

Active stereo. Depth may be determined by triangulating patterns of light projected onto a scene, instead of relying on the inherent texture of surfaces. Noise patterns may be projected to provide additional texture for passive stereo matching, though for high resolution capture of faces the advantage over [BBB*10] is unclear. [ZRY06] captures real-time face models using repeated sequences of codified structured light patterns. Despite compensating for subject motion, the authors note significant errors when the subject is speaking. Also, this approach has not been shown to achieve full facial coverage. For applications with little head motion, such systems may be sufficient, but comprehensive facial performance capture requires more freedom of movement.

Photometric stereo. Some geometric information (i.e. surface normals) can be inferred from measurements of a surface's reflectance function. Dense reflectance measurement [DHT*00,WMP*06] provides excellent reflectance information, but capture times are too high to be practical for dynamic scenes. Reducing the resolution of the measurement allows these techniques to be extended to dy-

namic scenes [HWT*04], but still requires a large number of photographs. Photometric stereo using three colored lights enables dynamic capture [Woo78, KHE10, HV10], but inhibits good reflectance capture and restricts the placement of cameras. Photometric stereo using gradient illumination encodes first order reflectance information, which enables recovery of realistic reflectance and surface normals when combined with polarimetry [MHP*07] or example-based analysis [CGD09], and can robustly handle dynamic performances [WGP*10]. Of these methods, gradient illumination stands out as an efficient and robust method for capturing detailed reflectance information without violating any of the criteria for comprehensive facial performance capture.

Shape template methods. Morphable models [BV99] allow an approximate face shape to be fitted to as few as one photograph, and have been applied towards facial reflectance estimation [FBLS05], but are too generic to recover highly detailed geometry and reflectance for a dynamic performance. For higher fidelity reconstruction of an individual, many works capture a high resolution static scan of the performer as a geometry template, and then track the template onto lower resolution dynamic data using optical flow techniques [BPL*03] or motion tracking markers [GGW*98, HWT*04, BLB*08]. Great results may be obtained (examples include The Matrix Reloaded, The Curious Case of Benjamin Button, The Digital Emily Project, Tron Legacy), at the expense of significant artistic effort, especially in facial regions where automatic methods fail to capture subtle nuances of a performance, most notably around the eyes and mouth. These limitations prevent us from considering template based methods for our system.

Hybrid techniques. [EVC07] obtains both scene geometry and reflectance from photographs taken under multiple views and multiple illumination conditions. The results are robust and detailed, but reflectance is assumed to be Lambertian, and it requires a large number of photographs which would be impractical for dynamic performances. Other works refine coarse geometry from one source with photometric normals from another source to provide high-frequency details [NRDR05, MJH*08, HW08, WGP*10], but the coarse geometry may have artifacts that are too strong to be removed by the refinement. At the other end of the spectrum, [VPB*09] uses surface normal integration to extract per-viewpoint geometry with good high frequency detail but low frequency distortion, and then merges the geometries from multiple viewpoints to reduce the distortions. However, the geometry obtained has not been shown to be detailed enough for photorealistic facial performances.

3. Comprehensive Facial Performance Capture

Based on our review of prior work, the most promising approaches for comprehensive facial performance capture are those that operate on data recorded from multiple views, with a small number of multiplexed view-agnostic illumination conditions, and require no shape template or static scans. We therefore borrow from previous works as follows: Like [GGW*98], we capture a dynamic facial performance from multiple cameras simultaneously. Like [WGP*10], we robustly align temporally multiplexed gradient illumination photographs to obtain photometric normals for every frame of a performance. Inspired by [CGD09], we estimate facial reflectance from gradient illumination photographs using a novel heuristic. Inspired by [SSZ02, SFVG04, EVC07, HW08], we derive a principled likelihood model combining multi-view stereo correspondence with photometric stereo. Inspired by multi-resolution graph cuts [HMJI09] and multi-resolution belief propagation [VTSC04, YWA10], we employ a novel multi-resolution optimization approach, interleaving discrete domain and continuous domain belief propagation, yielding a result free of quantization artifacts. Finally, inspired by [BBB*10] and others, we merge depth estimates from multiple viewpoints into a single geometry. The output mesh and reflectance maps are suitable for rendering in off-the-shelf software.

4. Data Acquisition

We acquire a sequence of photographs capturing an actor under seven different illumination patterns, from five different views, for every frame of a performance. The illumination patterns are produced by a programmable illumination device, with 600 LED lights arranged in a sphere, similar to [MHP*07]. We position five Phantom v640 high-speed cameras around the front of the sphere in an "M" configuration (see Figure 1) to provide adequate coverage of the actor's face with some freedom for head rotation and movement. The resolution of the cameras (1600×1936) is sufficient to image fine skin details including pores. We calibrate the cameras using the method in [Zha00]. The illumination device cycles through seven gradient illumination patterns, similar to [WGP*10], which we call x, y, z, $\bar{x}$, $\bar{y}$, $\bar{z}$, and w. The x, y and z patterns are linear gradients over the sphere from 0 brightness to full brightness, aligned to the world coordinate axes. The $\bar{x}$, $\bar{y}$ and $\bar{z}$ patterns are linear gradients aligned to oppose the x, y and z patterns. The w pattern is uniform half brightness over the entire sphere. Together, these patterns encode first-order information for the reflectance function at every point on a surface. We also include four additional illumination patterns for validation, for a total of eleven patterns. The cameras are synchronized to the illumination device to record an entire set of patterns for every frame of performance capture output. In our tests, the output rate is 24 fps, requiring 264 fps photography (or 168 fps without validation patterns). It is also possible to reduce the rate of photography with minor losses in quality as shown in [WGP*10], which may allow the use of lower-cost cameras for performances having little rapid movement. To improve actor comfort, we triple the rate of cycling through the illumination patterns, to eliminate any perceptible flicker. To record a complete set of triple-rate patterns without increas-

ing the rate of photography, we expose each photograph for only the first third of the shutter interval. The actor also wore dark contact lenses, since the light is relatively bright.

5. Reflectance Estimation

We estimate reflectance information independently for each camera, starting from gradient illumination photographs aligned using the technique of [WGP*10] to avoid artifacts from fast head motion. We perform heuristic diffuse / specular separation of the uniform illumination photograph I_w, and then find a surface normal that best explains the other gradient illumination photographs I_x, I_y, I_z. Our heuristic is based on the observation that a gradient illumination photograph with the bright pole of the gradient facing the camera qualitatively appears to lack Fresnel reflections on the face (see z in figure 2), so the opposing gradient photograph should somewhat resemble the specular component (see $\bar{z}$ in figure 2). As gradient illumination is *steerable*, we *computationally* rotate the illumination in the photographs to align to the camera. We estimate the specular component S as the minimum of the three color channels of this image, since specular reflectance is typically achromatic. We then subtract S from twice the uniform illumination photograph ($2I_w$) to obtain an estimate of the diffuse albedo D. Assuming a Lambertian diffuse lobe plus a specular lobe centered about the ideal specular reflection direction, the gradient illumination photograph with gradient axis β has intensity:

$$G(\beta) = \tfrac{1}{2}D(1 + k_D n \cdot \beta) + \tfrac{1}{2}S(1 + k_S r \cdot \beta), \qquad (1)$$

where n is the surface normal, $r = (2n \cdot v)n - v$ is the ideal specular reflection direction, v is the view direction, $k_D = \frac{2}{3}$ and $k_S \approx 1$ [MHP*07]. We next search for the surface normal n minimizing $|G(\hat{x}) - I_x|^2 + |G(\hat{y}) - I_y|^2 + |G(\hat{z}) - I_z|^2$. Due to the form of (1), we may assume n lies somewhere between α and $(\alpha + v)/\|\alpha + v\|$, reducing the problem to a one-dimensional search, where α is the centroid of the reflectance function ($\langle I_x - I_w, I_y - I_w, I_z - I_w \rangle$, normalized). After the surface normals are estimated for every pixel, we apply an unsharp mask sharpening filter to the surface normal map. We use a constant filter width of 5 pixels in all examples, chosen to approximately cancel the subsurface scattering properties of human skin (from [WMP*06]), considering the typical width of one pixel in real world units. Finally, we compute a *view-independent* specular albedo $S' = S/F(n, v)$, where $F(n, v)$ is a generic Fresnel term. Together, the diffuse albedo D, specular albedo S', and surface normal n parameterize our reflectance model. Figure 2 shows the input photographs, and the estimated reflectance parameters.

6. Geometry Estimation

We take a principled maximum likelihood approach to geometry estimation. Inspired by [SSZ02] and [SFVG04], we consider the following likelihood model:

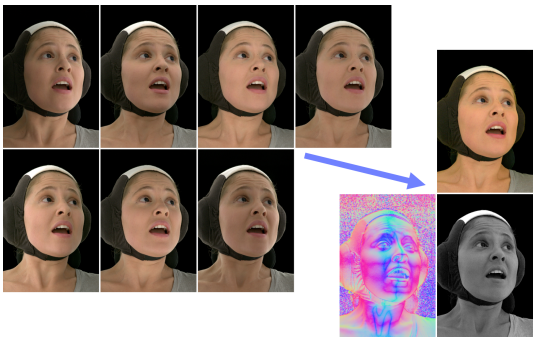

Figure 2: *Reflectance estimation. Upper left: x, y, z, w, $\bar{x}$, $\bar{y}$, $\bar{z}$ gradient illumination photographs. Lower right: estimated diffuse albedo, surface normal, and specular albedo.*

$$P(X, R, O|I) = \frac{P(I|X, R, O)P(X, R, O)}{P(I)} \qquad (2)$$

where X is a vertex position field, R is a reflectance function field, O is an occlusion state field, and I is the set of input images. We make a similar simplification to the occlusion term as [SSZ02] and assume O is independent of X, R, yielding:

$$P(X, R, O|I) \propto P(I|X, R, O)P(X|R)P(R)P(O) \qquad (3)$$

(It is also possible to iteratively update $P(O)$ as in [SLKS05], but we obtain acceptable results for faces using the simpler scheme.) We eliminate R, O by max marginalizing on X:

$$P(X|I) = \max_{R,O} P(X, R, O|I) \approx P(X, \bar{R}(X), \bar{O}(X)|I), \qquad (4)$$

where, ignoring $P(X|R)$ to make the solution tractable,

$$\bar{R}(X), \bar{O}(X) = \arg\max_{R,O} P(I|X, R, O)P(R)P(O). \qquad (5)$$

We are left with the following form:

$$P(X|I) \propto P(I|X, \bar{R}(X), \bar{O}(X))P(X|\bar{R}(X))P(\bar{R}(X))P(\bar{O}(X)) \qquad (6)$$

We do not model any spatial correlations in R or O, and we model pairwise spatial correlations in X, factoring (6) to:

$$P(X|I) \propto \prod_{s \in \mathbf{S}} P(I|x_s, \bar{r}_s(x_s), \bar{o}_s(x_s))P(\bar{r}_s(x_s))P(\bar{o}_s(x_s)) \prod_{(s,t) \in \mathbf{N}} P(x_s, x_t|\bar{r}_s(x_s), \bar{r}_t(x_t)) \qquad (7)$$

where $\mathbf{S}$ is the set of all sites in the field X, and $\mathbf{N}$ is the set of all ordered pairs of neighboring sites in the field X. The unit terms (in the product over $\mathbf{S}$) represent photometric consistency, reflectance likelihood, and occlusion likelihood. The pairwise terms (in the product over $\mathbf{N}$) represent a shape prior, trained on reflectance. At a high level, the model is related to the cost function in [EVC07] in that it contains a term for photometric consistency and a term for shape based on photometric normals, however theirs treats reflectance as Lambertian, using a threshold to remove specular highlights,

whereas our model estimates a more general reflectance distribution. Our model is also related to that in [HW08], but we do not approximate the photometric consistency term with a Gaussian distribution (which we believe poorly models sites with multiple high probability disparities), and we also model occlusion. In the following subsections, we detail the unit and pairwise terms.

6.1. Unit Term

The unit term $P(I|x_s,r_s,o_s)P(r_s)P(o_s)$ has three sub-terms. The first sub-term represents photometric consistency:

$$P(I|x_s,r_s,o_s) = \prod_{k \notin o_s} P(I_k(x_s)|r_s) \tag{8}$$

where $I_k(x_s)$ represents the pixel values in the photographs from camera k at projected position x_s, and o_s is the set of cameras for which x_s is occluded. Inspired by [SFVG04], we model $P(I_k(x_s)|r_s)$ as a normal distribution in terms of r_s:

$$P(I_k(x_s)|r_s) = \mathcal{N}(r_s;\mu_k(x_s),\Sigma_k(x_s)) \tag{9}$$

Where $\mathcal{N}(x;\mu,\Sigma)$ is the multidimensional normal distribution with mean vector μ and variance matrix Σ. In our work, r_s is a vector consisting of diffuse albedo, specular albedo, and surface normal. The mean vector $\mu_k(x_s)$ is heuristically estimated from $I_k(x_s)$ as in Section 5. Figure 3 shows the reflectance estimate for each of the five viewpoints in one frame of our tests. We use a camera sensor noise model to estimate the uncertainty in $I_k(x_s)$, and propagate the uncertainty through the computations in Section 5 using standard methods, yielding an estimate of the matrix $\Sigma_k(x_s)$. Besides camera noise, we also provide tunable parameters θ_i to model additional sources of uncertainty, modifying the diagonal of $\Sigma_k(x_s)$ by $\Sigma_k(x_s)_{i,i} \leftarrow (\Sigma_k(x_s)^{-1}_{i,i} + \theta_i^2)^{-1}$.

The second sub-term, $P(r_s)$, represents a priori reflectance likelihood. As r_s already parameterizes the reflectance model defined in Section 5, we simply ignore this term.

The third sub-term, $P(o_s)$, represents a priori occlusion likelihood. We model this as:

$$P(o_s) = P(\text{occ})^{\|o_s\|} \tag{10}$$

where $P(\text{occ})$ is the a priori likelihood of a pixel being occluded, provided as a tunable parameter. Despite its simplicity, this term is sufficient to suppress blob artifacts which would otherwise appear in occluded regions.

Since we do not model spatial correlations in R or O, we may compute $\bar{r}_s(x_s)$ and $\bar{o}_s(x_s)$ independently for each site:

$$\begin{aligned}\bar{r}_s(x_s),\bar{o}_s(x_s) &= \arg\max_{r_s,o_s} P(I|x_s,r_s,o_s)P(r_s)P(o_s) \\ &= \arg\max_{r_s,o_s} P(\text{occ})^{\|o_s\|} \prod_{k \notin o_s} \mathcal{N}(r_s;\mu_k(x_s),\Sigma_k(x_s)).\end{aligned} \tag{11}$$

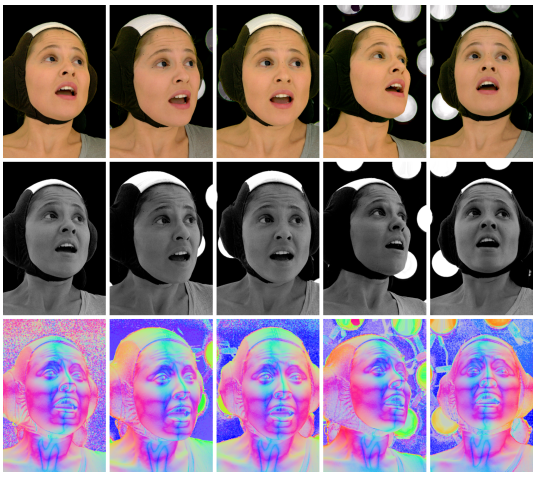

Figure 3: *Estimated reflectance parameters used as input to the geometry optimization. Top row: diffuse albedo. Middle row: specular albedo. Bottom row: surface normal. The first column is the center camera (the subject's head is turned somewhat to her left).*

For any assignment to o_s, we may maximize over r_s by:

$$\begin{aligned}r_s^\star(x_s,o_s) &= \arg\max_{r_s} \prod_{k \notin o_s} \mathcal{N}(r_s;\mu_k(x_s),\Sigma_k(x_s)) \\ &= \left[\sum_{k \notin o_s} \Sigma_k(x_s)^{-1}\right]^{-1} \left[\sum_{k \notin o_s} \Sigma_k(x_s)^{-1}\mu_k(x_s)\right]\end{aligned} \tag{12}$$

leaving:

$$\begin{aligned}\bar{o}_s(x_s) = \arg\max_{o_s} &P(\text{occ})^{\|o_s\|} \\ &\prod_{k \notin o_s} \mathcal{N}(r_s^\star(x_s,o_s);\mu_k(x_s),\Sigma_k(x_s))\end{aligned} \tag{13}$$

which we compute by exhaustive search over all possible assignments to o_s. And finally, $\bar{r}_s(x_s) = r_s^\star(x_s,\bar{o}_s(x_s))$.

6.2. Pairwise Term

The pairwise term $P(x_s,x_t|\bar{r}_s(x_s),\bar{r}_t(x_t))$ represents a shape prior. Motivated by [HMJI09] and [HW08], we incorporate photometric normals into the prior, instead of generic smoothing. To simplify optimization, we let the domain of X be a depth map from the view of a primary camera j. Then:

$$P(x_s,x_t|\bar{r}_s(x_s),\bar{r}_t(x_t)) = P(d_s,d_t|\bar{r}_s(x_s),\bar{r}_t(x_t)) \tag{14}$$

where $x_s = p + v_s d_s$, p is the nodal point of camera j in world space, v_s is the direction of the camera ray passing through the center of the pixel at site s, d_s is the depth at site s, and likewise for subscript t. It is possible to construct a likelihood model in terms of d_s,d_t for use in a belief propagation framework, including the effects of perspective and independent reflectance estimates for each depth. However, the mes-

sage passing step in belief propagation would be costly considering the entire domain of possible depth pairs, and we would rather take advantage of distance transforms [FH04] to efficiently compute messages. We may employ distance transforms if we express (14) in terms of the discrete depth gradient $d_t - d_s$ instead of d_s, d_t, ignoring the effects of perspective and removing the dependency of $\bar{r}_s(x_s), \bar{r}_t(x_t)$ on d_s, d_t through an approximation:

$$P(d_s, d_t | \bar{r}_s(x_s), \bar{r}_t(x_t)) \approx P(d_t - d_s | \mu_j(x_s), \mu_j(x_t)). \quad (15)$$

If the two pixel sites s and t lie on a continuous surface, we may compute an expected depth gradient $\Delta d_{s,t} \approx d_t - d_s$ as:

$$\Delta d_{s,t} = \hat{d}_t \left(\frac{1}{2} - \frac{1}{2} \frac{n_s \cdot v_t}{n_s \cdot v_s} \right) - \hat{d}_s \left(\frac{1}{2} - \frac{1}{2} \frac{n_t \cdot v_s}{n_t \cdot v_t} \right) \quad (16)$$

where n_s, n_t are the surface normals in $\mu_j(x_s), \mu_j(x_t)$ and $\hat{d}_s, \hat{d}_t$ are prior depth estimates (initially a typical depth, and in later iterations of our optimization, the depth estimates of the previous iteration). We may then use standard methods for uncertainty propagation to incorporate $\Delta d_{s,t}$, $\Sigma_j(x_s)$ and $\Sigma_j(x_t)$ into a Gaussian model for $d_t - d_s$. If, however, s and t straddle a surface discontinuity, then the depth gradient may take on much larger values than those implied by the surface normals. We provide for this case by truncating the quadratic term in the Gaussian distribution to be no more than some constant. We truncate the quadratic to one side of the mean only, being the side that is further from zero, since we presume the depth gradient of the closer surface to have greater magnitude than that of the farther surface at the point of discontinuity. We heuristically model the truncation constant in terms of $\mu_j(x_s), \mu_j(x_t)$, favoring continuous surfaces where the reflectance parameters at s and t are similar.

6.3. Optimization

We seek to find a depth map maximizing the likelihood defined by (7), which we treat as an optimization problem that seeks to maximize an objective function (or minimize its negative logarithm). Such problems are rampant in computer vision, and many methods for optimization have been proposed (for example, [TF03] compares graph cuts and belief propagation for stereo reconstruction). We desire continuous domain depth values, as opposed to discretely sampled depth values, so that the subtleties captured by the photometric surface normals may appear in the final geometry. [HW08] obtains continuous domain depth values by using Gaussian belief propagation, however we do not wish to approximate our photometric consistency term as a single Gaussian. Another option would be to use a fusion move optimization scheme [LRRB09], but this requires the generation of candidate solutions for fusion, which is an inexact science. Instead, we use discrete domain belief propagation *interleaved* with continuous domain (Gaussian) belief propagation, profiting from the benefits of both. Motivated by recent work in multi-resolution optimization schemes [VTSC04, HMJI09, YWA10], we improve performance by starting with a low resolution version of the problem, and doubling the resolution between each pass of interleaved discrete domain and continuous domain belief propagation, until the original resolution is reached. For both the discrete and continuous domain belief propagation, we choose Tree-Reweighted Sequential Belief Propagation (TRW-S) [Kol06] for its convergence properties and simple implementation.

Initialization. We begin with a coarsely spaced depth domain sampling (64 samples in the figures in this paper) on a downsampled input, downsampled just enough so that the window of depth samples covers the entirety of the desired performance volume.

Discrete phase. In our discrete phase, the domain for each pixel in the depth map is a window of equally spaced depth samples centered around the most likely depth from the previous resolution. Unlike [VTSC04, YWA10], we allow the depth window centers to take *fractional* values, obtained from the most likely depth values from the previous *continuous* phase. Since the previous continuous solution has only half the resolution of the current phase, we query it with bilinear interpolation, randomly jittering the pixel coordinate by up to plus or minus one half pixel to avoid grid artifacts. We iterate discrete domain TRW-S belief propagation for a number of iterations (10 in the figures in this paper).

Continuous phase. In our continuous phase, the domain for each pixel in the depth map is a Gaussian distribution. The pairwise terms in our objective function are truncated Gaussians, so we simply omit the truncation. However, the unit terms in our objective function are not Gaussian, but are well represented by discrete sampling. We therefore approximate each pixel's unit term by a Gaussian distribution, using a least squares fit to the discrete unit term weighted by the estimated a posteriori likelihood distribution (belief) from the previous discrete phase. This weighting adapts the Gaussian approximation of the unit term to the previous discrete solution. We iterate continuous domain TRW-S belief propagation for a larger number of iterations (100 in the figures in this paper) since it is fast.

Iteration. After the discrete and continuous phases are executed, we restore the input to the next larger resolution, narrow the spacing of the discrete depth sampling by half, and jump back to the discrete phase again. We continue iterating until finally the original high resolution input is processed. This approach combines the benefits of a discrete solution with the benefits of a continuous solution, without the drawback of oversimplifying the objective function.

Normal adjustment. Sometimes spatially correlated bias in the estimated surface normals will introduce small cracks or seams in the resulting geometry, due to the inability of our likelihood model to represent such correlations. As a work-around, we pause the multi-resolution scheme just before the final resolution is computed, and perform an adjustment on the low frequencies of the normals in a similar style

to [NRDR05], presuming that the depth estimate of the previous resolution is a good depth estimate.

6.4. Removing Uncertain Pixels

[SSZ02] uses the entropy of the inferred a posteriori likelihood distribution to estimate the uncertainty in the recovered depth. We adopt the same strategy and cut out uncertain pixels in the result using an entropy threshold. After removing uncertain pixels, small random patches of pixels may remain due to uncertainty in the entropy, so we segment the remaining pixels along steep depth gradients and cull away any segments smaller than a threshold number of pixels.

6.5. Merging

Although our geometry estimation operates on photographs from multiple cameras, the optimization domain is a depth map from the viewpoint of a single camera. Therefore we repeat the geometry estimation from the viewpoint of each camera in our system, and then merge the results into a single mesh, by projecting the vertex positions and reflectance parameters into a common cylindrical coordinate system (see figure 4). We weigh the contribution of each camera by the inverse projected area between neighboring vertices, and feather the weights to reduce seams. We then perform hole filling in the cylindrical domain.

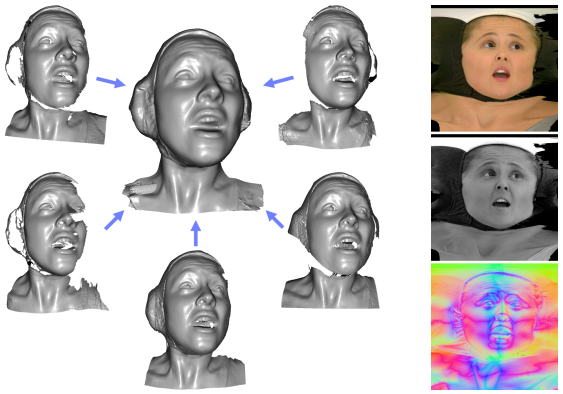

Figure 4: *Left: Depth maps computed from the viewpoint of each camera are merged into a single mesh. Right: Merged reflectance maps: diffuse, specular, normal.*

7. Results

We show results from our system for a live facial performance, processed automatically with no manual clean-up, and rendered under novel viewpoints and lighting conditions with subsurface scattering and global illumination using the V-Ray rendering software from Chaos Group. The running time of the optimization using a straightforward implementation on a single 3 GHz Intel processor core was 50 minutes per camera per frame, totalling 250 minutes per merged output frame. However, the different cameras and output frames may be processed in parallel on multi-core machines, so wall clock times were smaller. We also estimated per-vertex velocity using optical flow techniques, to enable physically-based motion blur rendering. Figure 5 shows renderings

Figure 5: *Illumination by a small spherical light source. Top row: ground truth. Bottom row: our result.*

under high-frequency illumination (a small spherical light source) compared to ground truth photographs, which were not used in the reconstruction. The estimated diffuse and specular albedos and surface normals are sufficient information to achieve photorealistic results even under these unforgiving illumination conditions. Figure 6 shows renderings

Figure 6: *Subsets of the captured information. Left to right: Raw geometry, addition of normal map, addition of diffuse and specular albedo maps.*

using the skin shader lit by a small spherical light source, using subsets of the captured information. The raw geometry contains enough detail to support gross shadowing and subsurface scattering effects. The normal map adds a life-like reconstruction of fine details like skin pores and blemishes even in the absence of any color texture. The diffuse and specular albedo maps provide the base color of the face and subtle variation in the highlights. Figure 7 shows selected renderings (every twentieth frame) from a fifteen second performance capture sequence, with novel camera mo-

Figure 7: *Every twentieth frame of a fifteen second performance, with novel camera motion, and motion blur.*

tion, and synthetic motion blur ($\frac{1}{96}$ second exposure). Using multiple cameras, short exposure photographs, and view-agnostic active illumination conditions allows the actor to deliver an emotional performance, including head rotation and rapid jerking motions, without compromising the fidelity of the reconstruction. Figure 8 illustrates the versatility attained by having highly detailed reflectance information, with renderings under image based lighting environments (from www.debevec.org). Photorealism is upheld under a wide variety of illumination conditions. Figure 9 shows a close-up rendering, highlighting the realism of subtle skin details under grazing lighting conditions. Facial performance capture systems that avoid active illumination simply do not achieve this level of photorealism. Figure 10 compares the proposed method to [WGP*10], for geometry obtained from two views. The quality of the geometry is comparable, with the proposed method recovering improved geometry around the sides of the face and neck, and notably the eyelids. Figure 11 shows the facial coverage afforded by our system. With five cameras, the entire face is captured even as the subject turns her head from 30° left to 30° right.

Figure 8: *Renderings under image based lighting environments. Photorealistic results are achieved under both soft and harsh direct lighting, including back lighting. (The black dots on the subject's face are not used by our system.)*

Figure 9: *The reflectance information captured by gradient illumination provides subtle skin details that remain lifelike even under grazing lighting conditions.*

8. Limitations and Future Work

While our system produces photorealistic reconstructions of facial performances, there are a number of approximations, shortcuts, or limitations we wish to address in the future. The most visually objectionable artifact is an occasional spike poking out of the face. The depth maps computed by our optimization exhibit no such artifacts, but sometimes have patches of background geometry floating beside the face that are not culled by our entropy thresholding. These floaters may be erroneously projected onto the face in the cylindrical domain during merging. We would like to use a better

Figure 10: *The proposed method compared to [WGP*10]. Left to right: Proposed (subject 1), [WGP*10] (subject 1), proposed (subject 2), [WGP*10] (subject 2).*

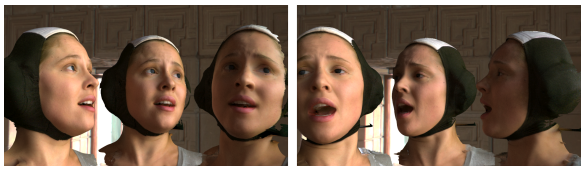

Figure 11: *Clones rendered from three viewpoints, for two frames of a performance, illustrating full facial coverage.*

merging algorithm, or to directly optimize a single geometry in a more suitable domain than a depth map. Initial tests using Poisson merging [KBH06] indicate that the spike artifacts can be avoided without relying on any post-processing. We may investigate better diffuse / specular separation to reduce the bias in the reflectance estimate. We may also seek to model the bias in the reflectance estimate, to reduce its effect on the reconstructed geometry, and to generally improve our principled likelihood model by making fewer approximations and assumptions. In our tests, the subjects captured wore a hat covering their hair and ears. An area wide open for future exploration is including hair and ears in the capture, possibly even extending all the way around the head using more cameras. Finally, methods for computing dense temporal correspondences across entire performances could be investigated, to facilitate any editing operations to be performed on the captured performances, including integration into conventional production pipelines.

9. Conclusion

In this work we present comprehensive facial performance capture. To our knowledge, this is the first facial performance capture system to achieve photorealistic reconstructions, with full facial coverage, of a dynamic, emotional performance under both novel viewpoints and novel illumination in an automatic setting. Facial features are reconstructed that are often omitted in previous work, such as eyes and teeth, though they could stand some improvement in future work. Fine surface details such as skin pores and blemishes are recovered faithfully. This level of realism is achieved by estimating detailed reflectance information and detailed geometry independently for each frame of the performance, without the need for any template geometry or static scans. This is made possible by two key contributions: The first is a novel heuristic for estimating detailed facial reflectance from gradient illumination photographs. The second is a novel geometry optimization framework that maximizes a principled likelihood function combining multi-view stereo correspondence and photometric stereo, using a novel multi-resolution belief propagation approach combining discrete domain and continuous domain belief propagation. We demonstrate the realism of performances captured by our system with photorealistic renderings made in off-the-shelf software, with subsurface scattering, global illumination, and motion blur. The renderings are a close match to ground truth photographs. Photorealism is maintained both under novel viewpoints and under novel illumination, including high-frequency illumination, back-lighting, and grazing illumination.

Acknowledgements

We wish to thank Randy Hill, Bill Swartout, Cheryl Birch, Hanna Dershowitz, Kathleen Haase, Clay Sparks, Jules Urbach, Alfonso Cuaron, Chris DeFaria, Mark Brown, Anna Panton, Chris Lawrence, Tim Webber, Nikki Penny, the reviewers for their time and comments, our subject for her emotional performance, our colleagues for proofreading, and Chaos Group for the V-Ray rendering software.

References

[ARL*09] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG M., DEBEVEC P.: Creating a photoreal digital actor: The digital emily project. *Conference for Visual Media Production* (2009), 176–187. 1

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-quality single-shot capture of facial geometry. In *SIGGRAPH '10: ACM SIGGRAPH 2010 papers* (New York, NY, USA, 2010), ACM, pp. 1–9. 2, 3

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. *ACM Trans. Graph. 29*, 4 (2010), 1–10. 1, 2

[BLB*08] BICKEL B., LANG M., BOTSCH M., OTADUY M. A., GROSS M.: Pose-space animation and transfer of facial details. In *SCA '08: Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2008), Eurographics Association, pp. 57–66. 1, 3

[BPL*03] BORSHUKOV G., PIPONI D., LARSEN O., LEWIS J. P., TEMPELAAR-LIETZ C.: Universal capture: image-based facial animation for "the matrix reloaded". In *ACM SIGGRAPH 2003 Sketches & Applications* (New York, NY, USA, 2003), SIGGRAPH '03, ACM, pp. 1–1. 1, 3

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1999), ACM Press/Addison-Wesley Publishing Co., pp. 187–194. 3

[CGD09] CHEN T., GHOSH A., DEBEVEC P.: Data-driven diffuse-specular separation of spherical gradient illumination. In *SIGGRAPH '09: Posters* (New York, NY, USA, 2009), SIGGRAPH '09, ACM, pp. 33:1–33:1. 3

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of ACM SIGGRAPH 2000* (July 2000), Computer Graphics Proceedings, Annual Conference Series, pp. 145–156. 2

[EVC07] ESTEBAN C. H., VOGIATZIS G., CIPOLLA R.: Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30* (2007), 548–554. 3, 4

[FBLS05] FUCHS M., BLANZ V., LENSCH H., SEIDEL H.-P.: Reflectance from images: A model-based approach for human faces. *IEEE Transactions on Visualization and Computer Graphics 11* (2005), 296–305. 3

[FH04] FELZENSZWALB P. F., HUTTENLOCHER D. P.: Efficient belief propagation for early vision. In *CVPR* (2004), pp. 261–268. 6

[FP09] FURUKAWA Y., PONCE J.: Accurate camera calibration from multi-view stereo and bundle adjustment. *IJCV 84*, 3 (September 2009). 2

[GGW*98] GUENTER B., GRIMM C., WOOD D., MALVAR H., PIGHIN F.: Making faces. In *SIGGRAPH* (1998), pp. 55–66. 1, 3

[GSC*07] GOESELE M., SNAVELY N., CURLESS B., HOPPE H., SEITZ S.: Multi-view stereo for community photo collections. In *ICCV07* (2007), pp. 1–8. 2

[HMJI09] HIGO N., MATSUSHITA Y., JOSHI N., IKEUCHI K.: A hand-held photometric stereo camera for 3-d modeling. In *ICCV09* (2009). 3, 5, 6

[HV10] HERNANDEZ C., VOGIATZIS G.: Self-calibrating a real-time monocular 3d facial capture system. In *3DPVT* (2010). 3

[HW08] HAINES T., WILSON R.: Combining shape-from-shading and stereo using gaussian-markov random fields. In *ICPR08* (2008), pp. 1–4. 3, 5, 6

[HWT*04] HAWKINS T., WENGER A., TCHOU C., GARDNER A., GÖRANSSON F., DEBEVEC P.: Animatable facial reflectance fields. In *Rendering Techniques* (2004), pp. 309–321. 1, 3

[KBH06] KAZHDAN M., BOLITHO M., HOPPE H.: Poisson surface reconstruction. In *SGP '06: Proceedings of the fourth Eurographics symposium on Geometry processing* (Aire-la-Ville, Switzerland, Switzerland, 2006), Eurographics Association, pp. 61–70. 9

[KHE10] KLAUDINY M., HILTON A., EDGE J.: High-detail 3d capture of facial performance. In *3DPVT* (2010). 3

[Kol06] KOLMOGOROV V.: Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell. 28*, 10 (2006), 1568–1583. 6

[LRRB09] LEMPITSKY V., ROTHER C., ROTH S., BLAKE A.: Fusion moves for markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 99*, PrePrints (2009). 6

[MHP*07] MA W.-C., HAWKINS T., PEERS P., CHABERT C.-F., WEISS M., DEBEVEC P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Rendering Techniques 2007: 18th Eurographics Symposium on Rendering* (June 2007), pp. 183–194. 3, 4

[MJH*08] MA W.-C., JONES A., HAWKINS T., CHIANG J.-Y., DEBEVEC P.: A high-resolution geometry capture system for facial performance. In *ACM SIGGRAPH 2008 talks* (New York, NY, USA, 2008), SIGGRAPH '08, ACM, pp. 3:1–3:1. 1, 3

[NRDR05] NEHAB D., RUSINKIEWICZ S., DAVIS J., RAMAMOORTHI R.: Efficiently Combining Positions and Normals for Precise 3D Geometry. *SIGGRAPH (ACM Transactions on Graphics) 24*, 3 (2005), 536–543. 3, 7

[SCD*06] SEITZ S., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR06* (2006), pp. I: 519–528. 2

[SFVG04] STRECHA C., FRANSENS R., VAN GOOL L.: Wide-baseline stereo from multiple views: A probabilistic account. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1* (jun. 2004), I–552 – I–559 Vol.1. 3, 4, 5

[SLKS05] SUN J., LI Y., KANG S., SHUM H.: Symmetric stereo matching for occlusion handling. In *CVPR05* (2005), pp. II: 399–406. 2, 4

[SS02] SCHARSTEIN D., SZELISKI R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV 47*, 1-3 (April 2002), 7–42. 2

[SSZ02] SUN J., SHUM H., ZHENG N.: Stereo matching using belief propagation. In *ECCV02* (2002), p. II: 510 ff. 3, 4, 7

[SZJ09] SMITH B., ZHANG L., JIN H.: Stereo matching with nonparametric smoothness priors in feature space. In *CVPR09* (2009), pp. 485–492. 2

[TF03] TAPPEN M., FREEMAN W.: Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV03* (2003), pp. 900–907. 6

[VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P., POPOVIĆ J., RUSINKIEWICZ S., MATUSIK W.: Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 28*, 5 (Dec. 2009). 3

[VTSC04] VOGIATZIS G., TORR P., SEITZ S., CIPOLLA R.: Reconstructing relief surfaces. In *BMVC* (2004), pp. 117–126. 3, 6

[WGP*10] WILSON C. A., GHOSH A., PEERS P., CHIANG J.-Y., BUSCH J., DEBEVEC P.: Temporal upsampling of performance geometry using photometric alignment. *ACM Trans. Graph. 29*, 2 (2010), 1–11. 1, 3, 4, 8, 9

[WGT*05] WENGER A., GARDNER A., TCHOU C., UNGER J., HAWKINS T., DEBEVEC P.: Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics 24*, 3 (Aug. 2005), 756–764. 1

[WMP*06] WEYRICH T., MATUSIK W., PFISTER H., BICKEL B., DONNER C., TU C., MCANDLESS J., LEE J., NGAN A., WANN H., GROSS J. M.: Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics 25* (2006), 1013–1024. 2, 4

[Woo78] WOODHAM R. J.: Photometric stereo: A reflectance map technique for determining surface orientation from image intensity. In *Proceedings of SPIE's 22nd Annual Technical Symposium* (Aug. 1978), vol. 155. 3

[WTRF09] WOODFORD O., TORR P., REID I., FITZGIBBON A.: Global stereo reconstruction under second-order smoothness priors. *PAMI 31*, 12 (December 2009), 2115–2128. 2

[YWA10] YANG Q., WANG L., AHUJA N.: A constant-space belief propagation algorithm for stereo matching. In *CVPR* (2010). 3, 6

[Zha00] ZHANG Z.: A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 11 (nov. 2000), 1330 – 1334. 3

[ZRY06] ZHANG S., ROYER D., YAU S.-T.: Gpu-assisted high-resolution, real-time 3-d shape measurement. *Opt Express 14*, 20 (2006), 9120–9. 2

[ZSCS04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S.: Spacetime faces: high resolution capture for modeling and animation. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers* (New York, NY, USA, 2004), ACM, pp. 548–558. 1