Temporal Upsampling of Performance Geometry Using Photometric Alignment

CYRUS A. WILSON, ABHIJEET GHOSH, PIETER PEERS, JEN-YUAN CHIANG, JAY BUSCH, and PAUL DEBEVEC
Institute for Creative Technologies, University of Southern California

We present a novel technique for acquiring detailed facial geometry of a dynamic performance using extended spherical gradient illumination. Key to our method is a new algorithm for *jointly* aligning two photographs, under a gradient illumination condition and its complement, to a full-on tracking frame, providing dense temporal correspondences under changing lighting conditions. We employ a two-step algorithm to reconstruct detailed geometry for *every* captured frame. In the first step, we coalesce information from the gradient illumination frames to the full-on tracking frame, and form a temporally aligned photometric normal map, which is subsequently combined with dense stereo correspondences yielding a detailed geometry. In a second step, we propagate the detailed geometry back to every captured instance guided by the previously computed dense correspondences. We demonstrate reconstructed dynamic facial geometry, captured using moderate to video rates of acquisition, for every captured frame.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*; I.4.1 [**Image Processing and Computer Vision**]: Digitization and Image Capture—*Imaging geometry*

General Terms: Measurement, Algorithms

Additional Key Words and Phrases: Capture, photorealism, 3D face scanning, motion estimation, optical flow

ACM Reference Format:

Wilson, C. A., Ghosh, A., Peers, P., Chiang, J.-Y., Busch, J., and Debevec, P. 2010. Temporal upsampling of performance geometry using photometric alignment. ACM Trans. Graph. 29, 2, Article 17 (March 2010), 11 pages. DOI = 10.1145/1731047.1731055 http://doi.acm.org/10.1145/1731047.1731055

1. INTRODUCTION

The creation of photo-real imagery of human faces has made significant strides forward in recent years, using increasingly detailed texture, geometric, and motion information. This article contributes to the acquisition and computation of high-quality dynamic facial geometry. Existing face scanning systems can acquire high-resolution facial textures and geometry, but typically only do so for static poses [XYZRGB ; Ma et al. 2007]. Maintaining a nonneutral expression for the duration of a scan is often difficult, and can introduce artifacts in the geometry reconstruction. There exist several real-time acquisition systems [Zhang et al. 2004; Zhang and Huang 2006; Ma et al. 2008]. However, these either sacrifice spatial resolution for real-time processing, or are data intensive and require high-speed photography equipment. The goal of this article is to produce detailed facial geometry of dynamic performances with the spatial fidelity of data-intensive methods, but with a minimal temporal data budget (i.e., without using high-speed cameras or projectors).

In this work, we present a novel technique for capturing and computing detailed dynamic facial geometry. We create dynamic high-resolution geometry by combining photometric normals, obtained from spherical gradient illumination [Ma et al. 2007], and dense stereo correspondences. Prior techniques that employ photometric stereo to capture performance geometry [Wenger et al. 2005; Malzbender et al. 2006; Ma et al. 2008] rely on temporal multiplexing of different lighting conditions, and consequently require an effective data rate that is a multiple of the final geometry rate. In this work, we reduce the effective data rate by (1) reducing the number of illumination conditions, and (2) exploiting temporal relations of frames under gradient illumination conditions and their complements. To reduce the number of illumination conditions, we obtain coarse base geometry from dense stereo correspondences at each full-on tracking frame *without* employing structured light patterns, unlike the approach of Ma et al. [2008]. We further exploit temporal relations using a novel robust photometric alignment algorithm, that jointly computes dense spatial correspondences from pairs of complementary gradient illumination frames to a uniform spherical

The work was supported by the U.S. Army Research, Development, and Engineering Command (RDE-COM) and the University of Southern California Office of the Provost. The content of the information does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

Authors' addresses: C. A. Wilson (corresponding author), A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec, Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Ray, CA 90292; email: cwilson@ict.usc.edu.

Fig. 1. Smiling sequence captured under extended spherical gradient illumination conditions (top row), synthesized intermediate photometric normals (center row), and high-resolution geometry (bottom row) of a facial performance as reconstructed with the same temporal resolution as the data capture.

illumination (full-on) frame that is employed as a tracking frame (Figure 3, red arrows). Once dense correspondences between every gradient pattern and its flanking tracking frames are computed, we employ a two-step algorithm to compute detailed facial geometry at every captured frame. In a first step, we coalesce gradient information to the full-on tracking frame to infer a high-resolution photometric normal map, and combine this with the coarse base geometry obtained from dense stereo correspondences. In a second step, we propagate the detailed geometry back to the individual gradient frames using the dense jointly computed correspondences (Figure 3, magenta arrows), effectively yielding a dynamic high-resolution geometry sequence at the same frame rate as the original data capture (see Figure 1).

The proposed extended gradient illumination conditions enable robust alignment for photometric stereo. This sequence of illumination conditions works in concert with our joint alignment approach, even though traditional warping algorithms, such as optical flow, would fail due to the changing illumination conditions. Furthermore, because the photometric normals are computed from a larger set of patterns (without any net temporal penalty), our estimates exhibit an improved signal-to-noise ratio. A direct application of the extended gradient illumination conditions, besides their use in temporal upsampling of dynamic facial performances, is the automatic correction of subject motion during a static geometry scan, in which geometric quality is of utmost importance.

In summary, the principal contributions of this work are:

(1) a new optical flow formulation which enables reliable alignment of images under different illumination conditions which satisfy a novel "complementation constraint"; and a new set of extended gradient illumination conditions which meet the constraint, allowing us to compensate for subject motion in order to robustly compute photometric normals;

(2) doubling of the signal-to-noise for the computation of photometric normals;

(3) exploiting the availability of motion-compensated high-quality photometric normals to augment coarse-resolution base geometry recovered without structured light patterns;

(4) a novel technique for temporal upsampling of the performance geometry, based on photometric alignment, to match the frame rate of the capture process.

2. RELATED WORK

3D face scanning. While there has been a wide body of work on 3D scanning of objects, we focus our discussion on scanning of human faces due to the specific challenges in obtaining high-quality geometry. There exist techniques for high-resolution scanning of static faces based on laser scanning of a plaster cast [XYZRGB]. However, such techniques are not well suited for scanning faces in nonneutral expressions. Several real-time 3D scanning systems exist that are able to capture dynamic facial performances. These methods either rely on structured light [Rusinkiewicz et al. 2002; Zhang et al. 2004; Davis et al. 2005; Zhang and Huang 2006], or use photometric stereo [Wenger et al. 2005; Malzbender et al. 2006; Hernandez et al. 2007]. Recently, photometric stereo has also been employed to obtain detailed dynamic full-body geometry [Ahmed et al. 2008; Vlasic et al. 2009]. However, these prior methods are limited: either they do not provide sufficient resolution to model facial details, they assume uniform albedo, or they are data intensive. Bickel et al. [2007] take an alternate approach by first acquiring a detailed static scan of the face including reflectance data, and then augmenting it with traditional marker-based facial motion-capture data for large-scale deformation, and two high-resolution video cameras for tracking medium-scale expression wrinkles. Ma et al. [2007] recently introduced a technique for high-resolution face scanning of

static expressions based on photometric surface normals computed from spherical gradient illumination patterns. These photometric normals are used to add fine-scale detail to base geometry (obtained from structured light) as in Nehab et al. [2005]. Ma et al. [2008] also extend the approach to capture dynamic performances by employing high-speed photography and time-multiplexed active illumination. In our work, we strive for the spatial scanning resolution similar to Ma et al. [2007], but further extend the scanning resolution in the temporal domain without resorting to data-intensive high-speed acquisition.

Motion estimation. Motion estimation and compensation techniques have been previously employed by real-time face scanning techniques to correct for subject motion during acquisition. In particular, optical flow algorithms have been used very successfully for alignment of tracking frames in high-speed acquisition [Wenger et al. 2005; Ma et al. 2008]. They have also been applied to temporally align differently exposed video frames for creation of high dynamic range video sequences [Kang et al. 2003], and for computing movement of 3D surface points in the context of human performances [Vedula et al. 2005]. In our work, we estimate motion between different gradient illumination frames for alignment of the photometric information. Motion estimation has been performed previously under varying distant illumination for rigid objects [Zhang et al. 2003]. However, we go further than previous work and successfully perform nonrigid motion estimation on images captured under different illumination conditions.

3. SYSTEM OVERVIEW

Our setup for capturing facial performances consists of an LED sphere with approximately 150 individually controllable lights that allow us to illuminate a subject with complex lighting conditions. We capture facial performances in two different settings. In the first setting, we employ a stereo pair of 10-Megapixel Canon 1D Mark III digital SLR cameras operating in "burst" mode, and which can capture up to 45 consecutive frames at 8fps rate. We demonstrate that even with this limited data rate, we can obtain high-quality 30fps temporally upsampled performance geometry for moderate motions. In a second setting, we employ a stereo pair of HD cameras (Vision Research Phantom V10s) that we operate at video rates of 30–60fps, allowing a wider range of natural performances to be captured.[1]

4. JOINT PHOTOMETRIC ALIGNMENT

We now describe our joint photometric alignment algorithm for estimating motion on images acquired under changing illumination conditions. We develop a joint optical flow algorithm for this purpose that specifically takes advantage of extended gradient illumination conditions for robust computation of motion under changing illumination. We express the joint optical flow error functional in the variational formulation of Brox et al. [2004]. However, the same ideas we describe in the following can also be expressed in other flow paradigms.

General optical flow. Optical flow is a 2D representation of the apparent motion of objects in an image. By assuming that a pixel maintains brightness (i.e., brightness constancy assumption) after motion, optical flow can be formally expressed as

$$I(\mathbf{x}, t) = I(\mathbf{x} + \mathbf{u}, t + 1), \tag{1}$$

where $I(\mathbf{x}, t)$ represents the pixel at location $\mathbf{x}$ and time t. Optical flow methods try to estimate the optical flow vector field $\mathbf{u}$ from a set of (temporally) ordered images by minimizing the error between the target image and the "flowed" source image:

$$\mathbf{u} \leftarrow \text{argmin}_{\mathbf{u}'} \, \varepsilon\left(I(\mathbf{x} + \mathbf{u}', t + 1), I(\mathbf{x}, t)\right), \tag{2}$$

where $\varepsilon(\cdot, \cdot)$ is the error metric employed by the optical flow system. However, this problem is ill-constrained in areas with low texture variation, and at occlusion boundaries. To make the problem tractable, a smoothness constraint is often added to exploit the fact that neighboring pixels are likely to move in a similar manner. There exist many optical flow methods that often use one or a combination of the following methods: phase correlation, block-based methods, differential methods (including variational methods), and discrete optimization methods (such as belief propagation). See Scharstein and Szeliski [2002] for an extensive evaluation of the quality and robustness of various optical flow algorithms.

In order to reliably obtain photometric normals during a performance, we need to compensate for subject motion under the different lighting conditions. This can be done by computing the optical flow between a set of captured gradient illumination images to a common tracking frame (under full-on illumination) and warping the former to the configuration of the tracking frame. However, due to the brightness constancy assumption of Eq. (1), optical flow algorithms cannot typically accommodate changes in illumination between images, because they attempt to attribute differences between images entirely to motion. Recent algorithms, such as, for example, Brox et al. [2004], have been designed to be more robust to moderate illumination changes. However, the gradient illumination conditions of Ma et al. [2007] are specifically designed to stimulate the different components of the surface normals as much as possible, and thus result in significant differences in brightness. Consequently the change in observed spatial variation in brightness over the subject is too strong for these algorithms to reliably compute the optical flow. Figure 2 shows the effects and the resulting normal map of warping a gradient lighting image to a full-on gradient image. As can be seen, the changes in brightness are too significant for the optical flow algorithm of Brox et al. [2004] to handle reliably. Other optical flow algorithms will yield results of similar quality.

Complementation constraint. In order to reliably obtain photometric normals during a performance, we need to compensate for subject motion. However, as noted before, conventional optical flow cannot typically accommodate significant changes in illumination. To extend optical flow for robustly aligning photometric information, we propose the following strategy: for each illumination condition we also capture the subject under a complementary illumination condition, such that the sum of an illumination image and its complement will be equal to the full-on illumination image, save for subject motion. From this we can formulate a "complementation constraint"

$$A(\mathbf{x}) = X(\mathbf{x} + \mathbf{u}) + \bar{X}(\mathbf{x} + \mathbf{v}). \tag{3}$$

The complementation constraint allows us to *jointly* align a gradient illumination image X and its complementary illumination image $\bar{X}$ to the full-on illumination image A using *two* warp functions $\mathbf{u}$ and $\mathbf{v}$. In our case, the complement to a spherical gradient illumination pattern of Ma et al. [2007] is simply the same pattern with the gradient direction reversed. The resulting extended gradient illumination conditions are repeated as a *set* of eight conditions

[1] While these cameras can capture at much higher frame-rates, we explicitely limit the acquisition rate to 30–60fps, a rate attainable by an increasing number of off-the-shelf cameras.

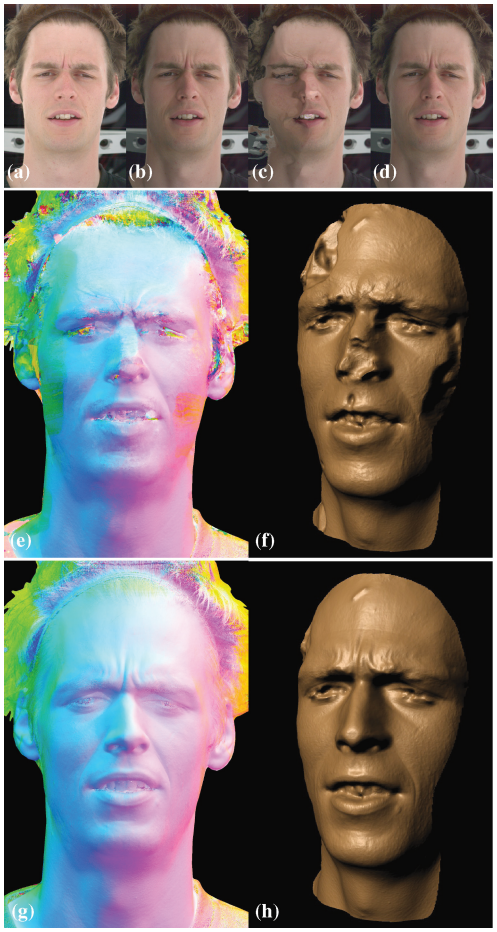

Fig. 2. Failure of conventional optical flow due to illumination change. (a): Tracking frame (under full-on illumination). (b): X gradient illumination frame. (c): X gradient image aligned to tracking frame using the optical flow algorithm of Brox et al. [2004]. (d): X gradient image aligned to tracking frame using joint photometric alignment with complementary image $\bar{X}$. (e,f): Normal map and reconstructed geometry using optical flow. (g,h): Using the proposed photometric alignment algorithm.

$\{X, Y, Z, A, \bar{X}, \bar{Y}, \bar{Z}, A\}$ during a facial performance capture. However, it should be noted that the complementation constraint is not specific to spherical gradient illumination and can be applied to any set of illumination condition pairs, as long as the sum of every pair corresponds to a common tracking frame.

Implementation. In our implementation, we base our joint photometric alignment on the variational optical flow framework of Brox et al. [2004]. While rewriting the error functionals in terms of the complementation constraint could yield a more efficient solver, it would also be very specific to the chosen optical flow algorithm. In anticipation of future advances in optical flow algorithms, we would like to use existing optical flow methods unaltered as a plug-in. To achieve this we exploit the fact that we can formulate an aligned gradient image as the difference between the full-on image and the reverse gradient image by employing the complementation constraint (Eq. (3)):

$$X(\mathbf{x}+\mathbf{u}) = A(\mathbf{x}) - \bar{X}(\mathbf{x}+\mathbf{v}). \tag{4}$$

A similar relation holds for $\bar{X}$. If one of the optical flows is known, then the difference image remains constant and can be directly plugged into an optical flow method to compute the other flow. However, since neither of the optical flows is known beforehand, we need to find a way to bootstrap the computations. We observe that even if the initial flow estimate is merely an approximation of the true flow, that a good approximation of the complementary flow can still be computed. This allows us to formulate an iterative algorithm where we alternate between solving for one of the two warps, while keeping the other fixed. The solution at iteration step $i+1$ can be computed by solving the minimization problems in the following using the solution from the ith step.

$$\begin{aligned} \mathbf{u}^{(i+1)} &\leftarrow \operatorname{argmin}_{\mathbf{u}} \varepsilon\big(X(\mathbf{x}+\mathbf{u}),\, A(\mathbf{x}) - \bar{X}(\mathbf{x}+\mathbf{v}^{(i)})\big) \\ \mathbf{v}^{(i+1)} &\leftarrow \operatorname{argmin}_{\mathbf{v}} \varepsilon\big(\bar{X}(\mathbf{x}+\mathbf{v}),\, A(\mathbf{x}) - X(\mathbf{x}+\mathbf{u}^{(i+1)})\big) \end{aligned} \tag{5}$$

Repeatedly applying this algorithm will converge to the correct solution (i.e., $\lim_{i\to\infty} \mathbf{u}^{(i)} = \mathbf{u}$, and, $\lim_{i\to\infty} \mathbf{v}^{(i)} = \mathbf{v}$).

While one could alternate after computing the optical flow completely before switching roles, it would take a considerable amount of time until convergence. For multi-resolution algorithms, one can alternate at every level, significantly improving performance. For iterative algorithms, one can alternate once every few iterations. In our implementation[2] we found that alternating every resolution level in the Brox et al. [2004] method yielded the convergence in the least amount of time.

To further improve the robustness of our flow computation, we also employ the "gradient constancy" assumption as described in Brox et al. [2004], by including the horizontal and vertical partial derivatives of the image pixel values as additional image channels. However, in this context it is difficult to determine a single global weighting factor to tune the relative influence of the derivative channels compared to the brightness channels, because the derivatives will vary spatially in scale according to the effects of the illumination condition. For example, the X gradient illumination condition will produce small derivatives on the left half of the face and large derivatives on the right half. Therefore we normalize the derivatives by dividing by a highly smoothed version of the image brightness.

5. TRACKING FRAME GEOMETRY

Using the available information from the stereo camera pair, dense stereo correspondences can be computed for the full-on tracking frames. Often due to lack of facial texture detail, the geometry can be of moderate quality. However, the preceding joint photometric alignment allows to warp the frames under extended gradient illumination to each full-on tracking frame, and compute a high-resolution motion-compensated photometric normal map. Using both the coarse base geometry and the high-resolution normal information, a detailed geometry can then be computed by embossing the photometric normals onto the coarse base geometry using the procedure of Nehab et al. [2005]. We now detail the computation of the photometric normals, and the coarse base geometry.

Photometric normals. We employ the joint alignment of Section 4 to align the temporally closest gradient illumination images and their complements to the target full-on tracking frame, generating a full set of six motion-compensated photometric images (see Figure 3). While it would be possible to directly apply the same

[2]We initialize both flows to zero at the coarsest level. At that scale this is a sufficiently good approximation to satisfy the complementation constraint and bootstrap the computations.

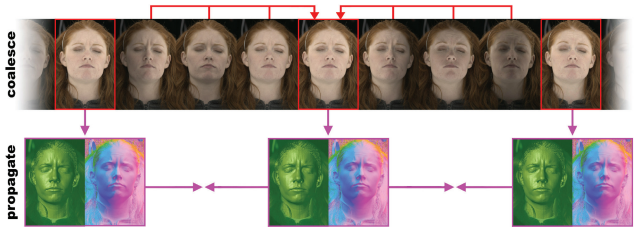

Fig. 3. An illustration of joint alignment and upsampling: Photometric normals are computed at tracking frames after joint alignment of gradient illumination conditions and their complements (red arrows). Base geometry is computed at the tracking frames from dense stereo correspondences. Both photometric normals and base geometries are then warped and blended at target intermediate frames (magenta arrows).

normal computation as in Ma et al. [2007], it would be a suboptimal use of the available information (only half of the photographs).

By exploiting the relation $X = -\bar{X}, Y = -\bar{Y}$, and $Z = -\bar{Z}$, we can write

$$\mathbf{n} = \frac{[X - \bar{X}, Y - \bar{Y}, Z - \bar{Z}]^{\mathrm{T}}}{\|[X - \bar{X}, Y - \bar{Y}, Z - \bar{Z}]^{\mathrm{T}}\|}. \tag{6}$$

This set of images are the same as $[X, Y, Z]$, and can be used instead to directly compute photometric normals. This improves the quality of the normal estimates, since pixels that are dark under one gradient illumination condition are most likely well exposed under the complementary gradient illumination condition. Note that even though we use twice the number of photographs for computing a normal map, no penalty on efficiency is incurred because we use each set of three gradient illumination conditions for the normal map computation at each of the two flanking tracking frames.

Base geometry. We also employ the Brox et al. [2004] optical flow algorithm employed in Section 4 to compute dense stereo correspondence between the left and the right camera pair at every tracking frame in order to recover base geometry. To improve the robustness of the stereo correspondences, we constrain the flow computations along the epipolar lines, and compute optical flow using both albedo as well as the aligned photometric normal information from the stereo camera pair. Note that any other algorithm that provides dense stereo correspondences can also be used [Scharstein and Szeliski 2002]. Furthermore, the accuracy requirements of the stereo correspondences are modest since most inaccuracies are compensated for when embossing the photometric normals. This allows us to compute coarse base geometry *without* using any additional information (such as additional structured light patterns) besides the extended spherical gradient illumination patterns.

6. INTERTRACKING FRAME GEOMETRY

The alignments computed in Section 4 describe a per-pixel warp from a nontracking capture frame (i.e., a gradient illumination image) to a tracking frame, and allows us to reconstruct detailed facial geometry (a combination of the coarse base geometry and photometric normals) at each tracking frame. These alignments can also be used to propagate the geometrical information back from the tracking frames to each individual gradient frame. This is attained in a straightforward manner by reversing their direction. One can employ these inverse flows to warp and blend together the normal maps and base geometries from the two tracking frames flanking the target frame.

Special care has to be taken when applying these inverse warps to handle occlusions at the eyelids and mouth correctly. An alternative solution is to compute new bidirectional flows between the tracking frames, and assume a linear deformation in between. To optimally use the available information, we also include normal information in addition to the albedo information at the tracking frames. However, we do not directly use the normal as a feature vector in the optical flow computation, because the orientation of a surface point can change between tracking frames. Instead, we use the partial derivatives of the x and y components of the photometric normal with respect to the corresponding image components u and v, that is, $\frac{\partial N_x}{\partial u}$ and $\frac{\partial N_y}{\partial v}$. This approach has the advantage that it deals better with occlusions, and allows to warp normal and geometrical information at a finer granularity than the capture rate. A disadvantage is the linear deformation assumption. However, this is usually not a major issue for moderate capture rates and regular facial motion, since tracking frames are only 4 frames apart.[3]

Once the correspondences are known between each tracking frame and each target frame, we can warp and blend both the base geometry and normal maps from the flanking tracking frames to the target frame. We call this procedure *temporal upsampling*. However, care has to be taken when blending the different sources of information. We start by explaining the basic warping on generic 2D images (i.e., albedo textures), and subsequently on geometries and normal maps.

Warping 2D images. Given the correspondences between a gradient frame and its flanking tracking frames, a warped 2D image is computed by blending pixels corresponding to the target pixel proportional to their temporal distance to the tracking frame.

Warping normal maps. A normal map is in essence a 2D map. However, care must be taken when blending the normals from flanking tracking frame normal maps. A simple linear blend would not result in a valid normal. Instead we rotate the normals proportional to the computed temporal weight and the angle between the normals at the tracking frames.

Warping base geometry. To warp the base geometry we exploit the fact that the base geometry obtained from stereo correspondences is essentially only 2.5D. This allows us to store each (x,y,z) coordinate at its 2D projected pixel location in one of the camera views. Warping is then similar to warping 2D images, where the color triplets have been replaced by the 3D coordinates. Note that this warp modifies 3D surface point locations, and not just depth (z) values.

Once both the base geometry and normal maps have been warped and blended from the flanking tracking frames, a high-resolution performance geometry[4] can be computed at the target intermediate time step by embossing the normals onto the geometry similar to the procedure of Nehab et al. [2005].

[3] Assuming linear subject motion during upsampling may result in the reconstructed motion deviating slightly from the true subject motion between tracking frames. However, it will still be accurate at tracking frames and temporally consistent across tracking frames. In contrast, assuming linear subject motion during *alignment* results in inaccuracies at tracking frames and artifacts which are temporally inconsistent across tracking frames.

[4] We do not perform explicit temporal smoothing of the geometry. However, since the process uses information coalesced from multiple captured gradient illumination frames, one might consider this an implicit temporal smoothing.

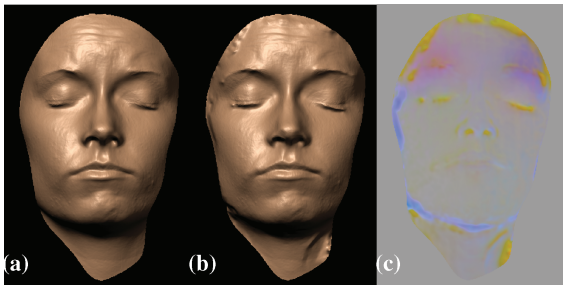

Fig. 4. Error analysis of the presented high-resolution geometry reconstruction algorithm compared to the technique of Ma et al. [2007]. (a) Using base geometry computed from stereo correspondence using structured light patterns. (b) Using base geometry computed in the absence of the structured light patterns. (c) Plot of difference in the reconstructed final high-resolution geometry when using (b) compared to (a), with contrast enhancement for illustration.

7. RESULTS

In this section, we present results of the temporal upsampling procedure for captured facial performances. As discussed in Section 3, we capture facial performance in two different settings. In the first setting, we capture performances with moderate amounts of motion at 8fps using digital SLR cameras in burst-mode for the duration of about 45 frames (after which the internal buffer of the camera fills and the frame rate drops significantly). Within this capture rate, full-on tracking frames are captured at a rate of 2fps. By applying the presented photometric alignment algorithm, we obtain a normal map and base geometry at every tracking frame. This is then further upsampled to the 4× upsampled final target frame rate of 32fps (see the accompanying video accessible at the ACM Digital Library). In a second setting, we capture performances with natural facial motion at regular video rates (30 to 60fps) using digital HD video cameras. Full-on tracking frames are captured at 7.5 to 15fps in this setting and the performances are then reconstructed after photometric alignment and temporal upsampling at the rate of the original video capture.

Figure 1 shows the various stages of the reconstruction of facial performance geometry starting with the set of proposed extended spherical gradient illumination conditions (top row) used to compute photometric normals (center row), computed using the alignment algorithm presented in Section 4. The motion estimated from the acquired photometric information is then used to warp the base geometry obtained from stereo correspondence and the photometric normals computed at the tracking frames to the depicted intermediate time steps. Final high-resolution geometry (bottom row) is then created at the depicted time steps by embossing the corresponding photometric normals onto the warped base geometry.

We provide an error analysis validation of our geometry reconstruction technique in Figure 4. In order to evaluate the robustness of our method, we compare the final high-resolution geometry reconstructed for a nonneutral expression in the presence (a) as well as absence (b) of the structured light scans for base geometry. Note that Figure 4(a) corresponds to the geometry reconstruction technique of Ma et al. [2007], while 4(b) corresponds to our strategy of obtaining dense stereo correspondence from optical flow constrained to epipolar lines between the cameras. As seen from the difference image 4(c), the deviation in the final reconstruction without using the structured light patterns for base geometry is very small (RMSE = 0.0678), supporting the thesis that any minor inaccuracies in the coarse detail of the base geometry are mostly compensated for by

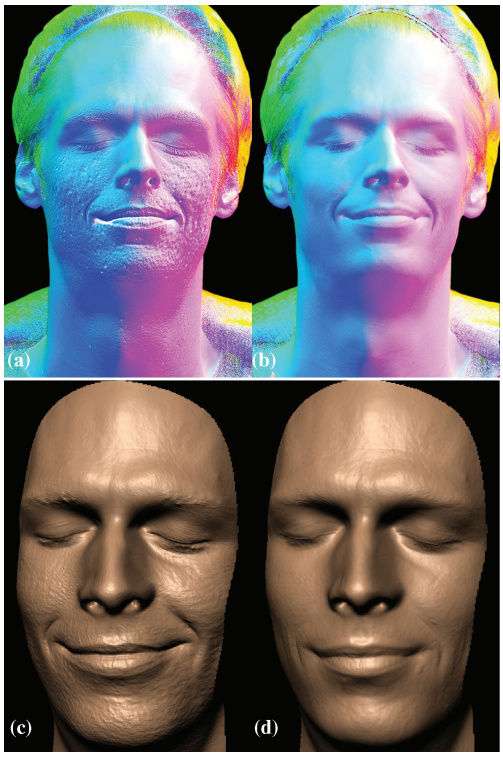

Fig. 5. Qualitative comparison of reconstruction of a scanned nonneutral expression with minor subject motion. (a,c): Reconstruction without alignment of the photometric information. (b,d): Reconstruction with the proposed photometric alignment algorithm. (a,b): recovered photometric normals. (c,d): reconstructed high-resolution geometry.

availability of the high-resolution photometric normals during the final embossing step. A similar observation was also made by Lim et al. [2005]. Note that the two techniques behave differently with respect to certain types of surface detail, such as hair. This can be seen in Figure 4 towards the edges of the geometry, where the differences in the error images are largest.

Figure 9 presents several examples of facial performances captured and reconstructed by our technique. The first presented example (top row) is reconstruction of a moderate facial motion using the digital SLRs at 8fps. This example demonstrates the robustness of our photometric alignment technique in the presence of significant facial motion despite the relatively low capture rate of the digital SLRs. The next two presented results (rows two and three) are examples of reconstructed high-resolution geometries of natural facial performances captured at regular video rates of 30fps (second row), and 60fps (third row) respectively using a pair of HD video cameras (2K resolution). Note how the presented technique is able to successfully reconstruct complex natural facial performances including eye and mouth motion *without* resorting to high-speed acquisition.

Applications. Figure 5 presents another qualitative benefit of our method for high-resolution scanning of a static nonneutral expression. Subject motion can occur even during scanning of static expressions, particularly for nonneutral expressions, requiring motion compensation for high-quality scan reconstruction. In this example, there was minor subject motion over the course of the scan resulting in motion artifacts and noise in the photometric normals Figure 5(a) obtained from the gradient illumination patterns with no alignment performed on the images. Our presented alignment

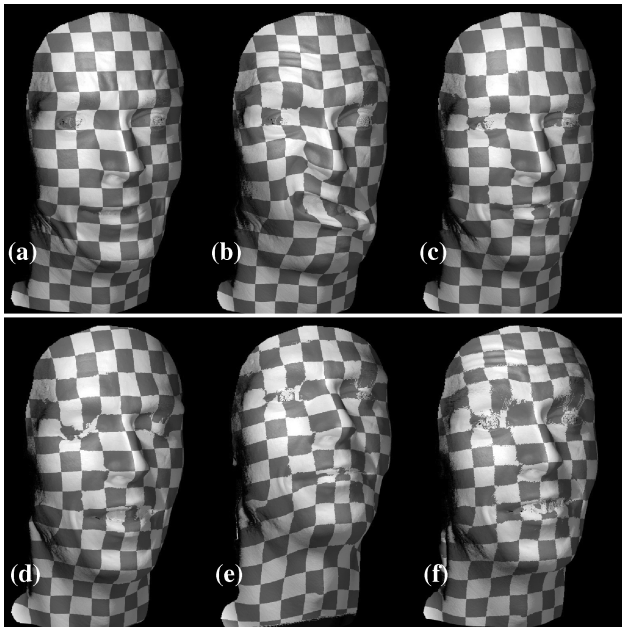

Fig. 6. Consistent u-v parameterization across a captured performance using photometric alignment. Here the u-v coordinates are warped according to the motion estimates from the acquired photometric information and used to texture map the reconstructed performance geometries ((a)-(f)) with an albedo map corresponding to the initial pose (a). Here we show a selection of 6 geometries over a sequence of 600 frames.

algorithm successfully computes the motion during the scanning process and compensates for it, producing a smooth photometric normal map free of artifacts Figure 5(b). The final high-resolution geometries obtained from both techniques are presented in the bottom row Figure 5(c) and 5(d). As seen in Figure 5, our photometric alignment algorithm produces a higher-quality artifact-free reconstruction 5(d) for such a nonneutral expression.

Figure 6 presents another application of our photometric alignment method for obtaining consistent u-v parameterizations between tracking frames of a captured facial performance. Here, we warp the u-v coordinates of the tracking frame base geometries with the warps obtained from the photometric alignment process to obtain temporally consistent u-v coordinates at intermediate time steps. Note that this application does *not* require that a subject be captured with mocap markers (in contrast to Ma et al. [2008], which placed dots on the subject's face). Similarly to accumulating optical flows over a sequence of images, obtaining a consistent mapping over multiple frames requires concatenating multiple warps. As such, it also suffers from the same limitation as concatenated optical flows, in particular a gradual drift from the correct correspondence. Note that the degree of drift depends only on the number of tracking frames, and not the number of upsampled frames. In practice we find that this simple approach of only considering motions between consecutive tracking frames is able to propagate a consistent u-v over distances of approximately 200 *tracking* frames, beyond which the correspondence degrades. In the accompanying video the quality of the correspondences over a 600 frame sequence, with significant subject movement, is illustrated using a checkerboard texture. The number of tracking frames (150) is comparable to the length of the performance sequences with dense u-v correspondence presented in Wand et al. [2009].

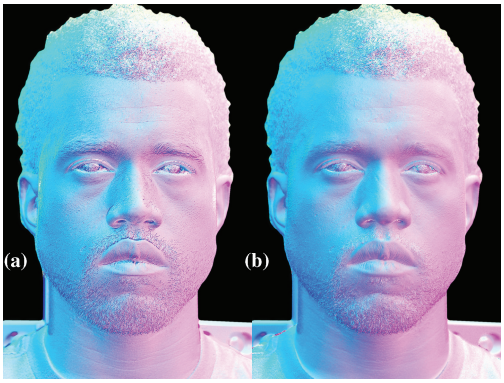

Fig. 7. Comparison of specular normals. (a) Specular normals obtained with the procedure of Ma et al. [2007]. (b) Specular normals obtained with a combination of parallel polarization and relighting of cross polarized diffuse albedo.

Specular detail. Ma et al. [2007] demonstrated that by separating specular reflections, so-called *specular normals* can be computed. Specular reflections are, in the case of skin, first surface interactions, and are not diluted by subsurface scattering as the diffuse reflections, and thus result in more accurate photometric normals. They compute separated specular reflections using polarization difference images under spherical gradient illumination. Extending this to time-multiplex illumination with alternating polarization for dynamic performances is technically difficult, since it would require high-speed photography in order to minimize subject motion between measurements under different polarization orientations.

The photometric alignment algorithm as discussed in Section 4 does not employ high-speed photography, and thus infers photometric information without polarization. As such, the fine-scale surface detail is obtained from less-than-ideal observations containing both diffuse and specular reflections. In order to improve the detail in the photometric normals, we have experimented with an alternative method for extracting specular detail. Figure 9 (bottom row) shows an example of a facial performance geometry reconstructed with *specular* photometric normals obtained with this technique. Instead of explicitly measuring the extended gradient illumination conditions under both parallel and cross polarization states, this alternative method only captures cross and parallel polarized imagery for the uniform spherical illumination condition, providing a direct measurement of the diffuse albedo, while capturing the remaining extended spherical gradient conditions only under parallel polarization (i.e., containing diffuse and specular reflections). We align the cross polarized photograph (containing the diffuse albedo) to the parallel polarized full-on tracking frame. Finally, we relight the observation of the diffuse albedo to each of the extended gradient conditions and subtract them from the aligned corresponding parallel polarized images to obtain an improved estimate of the specular component under spherical gradient illumination. This relighting process requires knowledge of the diffuse photometric normals, which we approximate from the parallel polarized images under spherical gradient illumination.

With this procedure, we are able to estimate approximate specular normals having taken only *one* additional measurement when using polarization compared to unpolarized measurements. The specular normals obtained with this procedure are a good approximation to those obtained with the procedure of Ma et al. [2007] (see Figure 7), while requiring fewer lighting conditions during acquisition. The minor differences that do arise are mainly caused by the difference

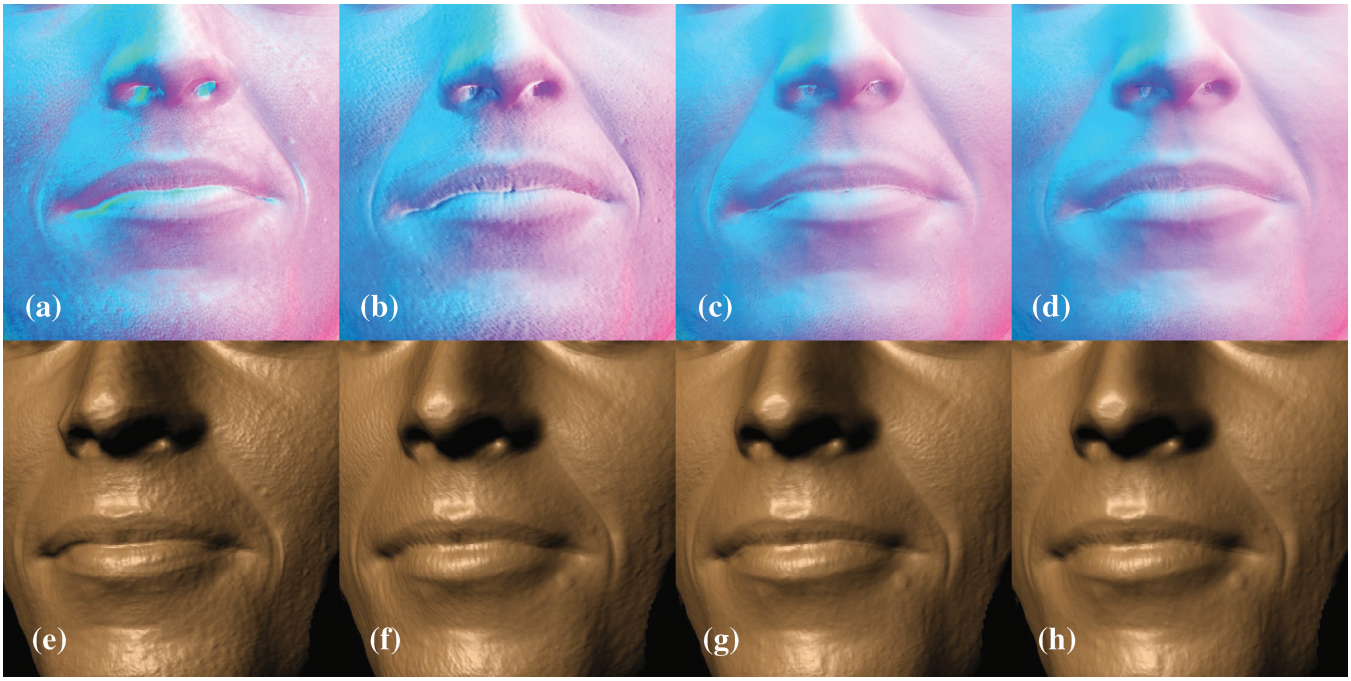

Fig. 8. Comparison with the optical flow formulation of Ma et al. [2008]. (a, b): Normal maps computed from gradient images aligned by assuming linear motion between tracking frames, using capture data at 30fps (a) and 60fps (b). Misregistration of the surface normal components leads to temporally inconsistent, spurious fine detail in the final geometry (e,f, video). Aligning each gradient image to the tracking frame using the proposed photometric alignment produces improved registration of normal map components (c,d) even at moderate frame rates such as 30fps (c) and 60fps (d), and temporally consistent fine detail (g,h, video).

in the estimated diffuse normals (obtained from parallel polarized data) used for relighting the diffuse albedo and the true photometric diffuse normals obtained with cross polarization.

8. DISCUSSION

Comparison with Ma et al. [2008]. While the overall goal of the proposed method (dynamic facial geometry capture) is similar to that of Ma et al. [2008], the underlying philosophy is significantly different. Ma et al. [2008] infer a single mesh for every 12 captured frames under different lighting conditions (7 structured light and 5 gradient patterns) by multiplexing, effectively requiring a 30×12 capture rate for a target 30fps geometry performance. In contrast, the proposed method is capable of generating a mesh for every captured frame. Visually, the quality of the results is very similar even though we only capture one *twelfth* of the data.

Ma et al. [2008] also propose to use optical flow between full-on tracking frames to compensate for any subject motion over the gradient illumination patterns. However, such motion compensation is suboptimal as it assumes linear motion between the full-on tracking frames, resulting in visual artifacts. A direct comparison between the proposed joint alignment and the alignment of Ma et al. [2008] is difficult because of the inclusion of the structured light patterns. Directly aligning complementary gradients over the structured light patterns impacts the quality of the alignment, resulting in an unfair comparison. A better comparison is to omit the structured light patterns, and compute the base geometry using stereo correspondences. This enables more methodical study of the effects of misalignments on the high-resolution geometry. Figure 8 (and the accompanying video) shows a comparison of the linear motion assumption between tracking frames employed by Ma et al. [2008] and the presented alignment method. Note that since both methods share the same base geometry, low- and medium-frequency geometrical details will be similar, and only the high-frequency content will differ. As can be seen in Figure 8 and the accompanying video, misalignment due to the linear deformation assumption gives rise to artifacts in the form of apparent additional surface detail. These spurious details appear and disappear over time: not with fine-scale deformations of the skin but with large-scale rigid movements of the head. The surface detail produced by joint photometric alignment, however, is more temporally coherent. Note that we compare both methods for video frame rates of 30 and 60fps.[5] At higher capture rates, the differences between both methods are less pronounced. However, using higher capture rates incurs a higher temporal budget and requires specialized high-speed equipment.

Computational cost. Note that while our approach enables capture at lower frame rates compared to previous methods, this comes at a significantly increased per-frame computational cost, considering the complicated optical flow computations, warping, etc., on high-definition images, from two cameras. With our current unoptimized implementation, the total computation time per frame (from raw input to final geometry) is 1 hour on a single 2.66 GHz CPU. It is

[5]Direct comparison requires that the subject's performance be identical between the 30fps and 60fps capture data. For this specific comparison we therefore prepared the 30fps input by skipping half the frames of the 60fps capture data.

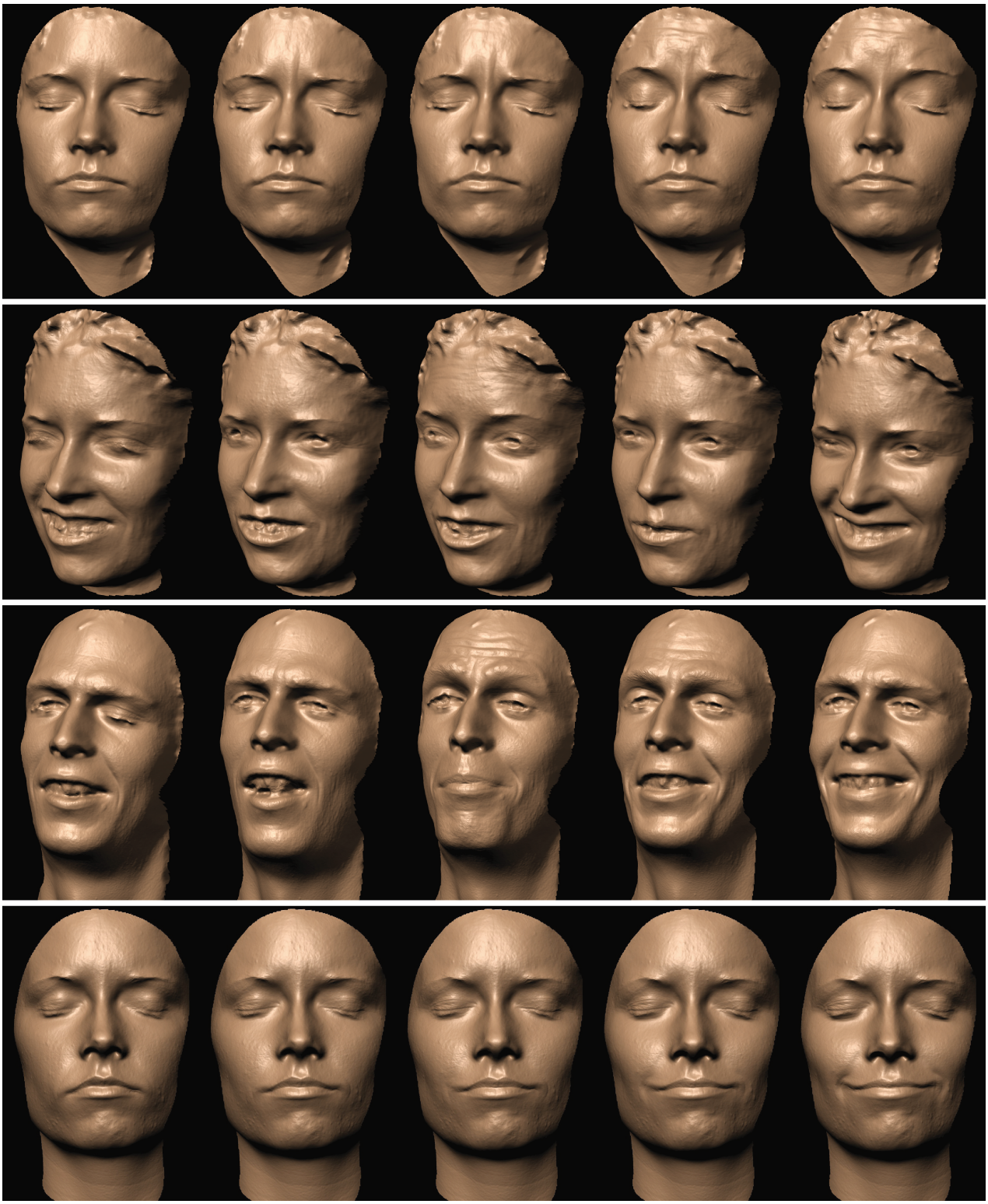

Fig. 9. Left to right: Various time steps of captured facial performance reconstructed by our proposed photometric alignment algorithm. Top row: An example of a performance with moderate motion captured at 8fps. Second row: Reconstructed geometry of a natural performance of a female subject captured at 30 fps. Third row: Reconstructed performance geometry of a male subject captured at 60 fps. Bottom row: An example of a captured facial performance reconstructed with specular normals.

worth noting that the cost of a joint photometric alignment is roughly double the cost of a single optical flow computation. Since we compute two warp functions for each joint alignment, no additional cost is incurred. Note also that multiple optical flow computations can in principle be run in parallel. We believe that the increased computational cost of the approach is worth the reduced capture frame rate requirement, as fast computers are significantly less expensive than specialized high-speed cameras; and the post-processing load can be distributed, whereas at the time of capture one is limited by camera storage capacity and physical limits such as available light levels.

Limitations. The presented photometric alignment technique assumes some smoothness in the motion; and the temporal upsampling step currently employs piece-wise linear interpolation. Reusing the jointly computed alignments suffers from occlusion artifacts. A combination of both techniques could potentially yield a more robust intertracking frame warping. Furthermore, the alignment algorithm currently does not handle changes in topology of the underlying geometry and assumes validity of photometric information everywhere in the captured photographs. Both these assumptions can be potentially violated in facial performances in regions inside the mouth, for example. Finally, the u-v parameterizations obtained from the alignment procedure may exhibit drift over extended performances due to the accumulation of small errors in optical flow computations between tracking frames.

9. CONCLUSION

We present a novel technique for temporal upsampling of captured facial performance geometry based on photometric alignment that enables us to reconstruct high-resolution photometric normals and geometry for every *captured* frame. We demonstrate several applications of such an alignment technique for both performance geometry capture and static scans as well as for obtaining u-v correspondences across a short- to moderate-length captured sequence. We introduce a novel joint alignment algorithm for computing motion in the presence of changing illumination, and demonstrate that existing optical flow frameworks can be easily modified to implement the proposed "complementation constraint." This complementation constraint is not specific to the spherical gradient illumination patterns, and potentially could be applied to other setups with appropriately designed multiplexed illumination.

For future work, it would be valuable to extend the approach to robustly handle changes in topology for more reliably reconstructing arbitrary facial performances, including speech and eye motion. It would also be advantageous to integrate the joint alignment implementation with more recent advances in optical flow techniques for potential performance benefits.

ACKNOWLEDGMENTS

We are extremely grateful to face-scanning subjects M. Lång, X. Yu, and K. West. We thank G. Stratou, A. Jones, and C. Milk for assistance with the capture process. We would also like to thank S. Mordijck, M. Nichelson, B. Swartout, R. Hill, and R. Hall for their generous support. Finally, we thank the anonymous reviewers for helpful suggestions and comments.

REFERENCES

AHMED, N., THEOBALT, C., DOBREV, P., SEIDEL, H.-P., AND THRUN, S. 2008. Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. 1–8.

BICKEL, B., BOTSCH, M., ANGST, R., MATUSIK, W., OTADUY, M., PFISTER, H., AND GROSS, M. 2007. Multi-Scale capture of facial geometry and motion. *ACM Trans. Graph. 26*, 3, 33: 1–10.

BROX, T., BRUHN, A., PAPENBERG, N., AND WEICKERT, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision*. 25–36.

DAVIS, J., NEHAB, D., RAMAMOORTHI, R., AND RUSINKIEWICZ, S. 2005. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Trans. Patt. Anal. Mach. Intell. 27*, 2, 296–302.

HERNANDEZ, C., VOGIATZIS, G., BROSTOW, G. J., STENGER, B., AND CIPOLLA, R. 2007. Non-Rigid photometric stereo with colored lights. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–8.

KANG, S., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2003. High dynamic range video. *ACM Trans. Graph. 22*, 3, 319–325.

LIM, J., HO, J., YANG, M.-H., AND KRIEGMAN, D. 2005. Passive photometric stereo from motion. In *Proceedings of the IEEE International Conference on Computer Vision*. 1635–1642.

MA, W.-C., HAWKINS, T., PEERS, P., CHABERT, C.-F., WEISS, M., AND DEBEVEC, P. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the Eurographics Symposium on Rendering*. 183–194.

MA, W.-C., JONES, A., CHIANG, J.-Y., HAWKINS, T., FREDERIKSEN, S., PEERS, P., VUKOVIC, M., OUHYOUNG, M., AND DEBEVEC, P. 2008. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Trans. Graph. 27*, 5, 121: 1–10.

MALZBENDER, T., WILBURN, B., GELB, D., AND AMBRISCO, B. 2006. Surface enhancement using real-time photometric stereo and reflectance transformation. In *Proceedings of the Eurographics Symposium on Rendering*. 245–250.

NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. 2005. Efficiently combining positions and normals for precise 3D geometry. *ACM Trans. Graph. 24*, 3, 536–543.

RUSINKIEWICZ, S., HALL-HOLT, O., AND LEVOY, M. 2002. Real-time 3d model acquisition. *ACM Trans. Graph. 21*, 3, 438–446.

SCHARSTEIN, D. AND SZELISKI, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision 47*, 1–3, 7–42.

VEDULA, S., BAKER, S., AND KANADE, T. 2005. Image based spatio-temporal modeling and view interpolation of dynamic events. *ACM Trans. Graph. 24*, 2, 240–261.

VLASIC, D., PEERS, P., BARAN, I., DEBEVEC, P., POPOVIĆ, J., RUSINKIEWICZ, S., AND MATUSIK, W. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph. 28*, 5, 174: 1–11.

WAND, M., ADAMS, B., OVSJANIKOV, M., BERNER, A., BOKELOH, M., JENKE, P., GUIBAS, L., SEIDEL, H.-P., AND SCHILLING, A. 2009. Efficient reconstruction of nonrigid shape and motion from real-time 3D scanner data. *ACM Trans. Graph. 28*, 2, 15: 1–15.

WENGER, A., GARDNER, A., TCHOU, C., UNGER, J., HAWKINS, T., AND DEBEVEC, P. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Trans. Graph. 24*, 3, 756–764.

XYZRGB. 3D laser scanning—XYZ RGB Inc. http://www.xyzrgb.com/.

ZHANG, S., AND HUANG, P. 2006. High-Resolution, real-time three-dimensional shape measurement. *Optical Engin. 45*, 12, 123601: 1–8.

ZHANG, L., CURLESS, B., HERTZMANN, A., AND SEITZ, S. M. 2003. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–625.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Space-time faces: High resolution capture for modeling and animation. *ACM Trans. Graph. 23*, 3, 548–558.

Received November 2009; accepted February 2010