Head-mounted Photometric Stereo for Performance Capture

Andrew Jones Graham Fyffe Xueming Yu Wan-Chun Ma Jay Busch Ryosuke Ichikari Mark Bolas Paul Debevec

University of Southern California Institute for Creative Technologies
jones@ict.usc.edu

Abstract

Head-mounted cameras are an increasingly important tool for capturing facial performances to drive virtual characters. They provide a fixed, unoccluded view of the face, useful for observing motion capture dots or as input to video analysis. However, the 2D imagery captured with these systems is typically affected by ambient light and generally fails to record subtle 3D shape changes as the face performs. We have developed a system that augments a head-mounted camera with LED-based photometric stereo. The system allows observation of the face independent of the ambient light and generates per-pixel surface normals so that the performance is recorded dynamically in 3D. The resulting data can be used for facial relighting or as better input to machine learning algorithms for driving an animated face.

1 Introduction

Realistic facial animation remains a major challenge in computer graphics as humans brains are wired to detect many different attributes of facial identity, expression, and motion. Advances in 3D scanning have enabled rapid capture of high-quality dense facial geometric and reflectance models that match real human subjects. This has led to many examples of compelling static virtual faces. The problem complexity, however, dramatically increases for believable facial motion.

Dynamic 3D scanning techniques typically require specialized cameras and projectors aimed at the face. The fixed hardware defines a limited capture volume so the subject's head must remain relatively stationary throughout the performance. Yet facial animation does not exist in a vacuum. Facial actions are often accompanied by full body actions. For example, eye-gaze follows the larger motion of the neck and torso and dialog is often accompanied by multiple hand gestures.

An alternative approach is to capture only sparse motion points using marker-based motion capture. Motion capture stages can accommodate multiple full-body performances and can scale up with additional cameras. Marker-based systems work well for bodies as markers can be placed at key joints to capture most degrees of freedom. Unfortunately, faces exhibit a significantly wider range of deformation that can not easily be represented by a simple set of bones and joints. Typically, faces require a dedicated set of up to 100 markers [26]. Even then, systems fail to capture important details around the mouth and eyes where it is not possible to place dense markers. Recently, commercial productions have started to use head-mounted cameras in motion capture environments to more accurately record dense sets of facial motion capture markers. These cameras have the advantages of providing a fixed video of the face even as the actor moves through a larger capture volume.

In this work we combine the accuracy of 3D scanning with the portability of head-mounted camera systems. Our key insight is that head-mounted cameras upgraded with active LED-based illumination can provide a richer range of dense facial geometric and motion cues with minimal increase in system weight. With two sequential photographs, we can capture video that is independent of ambient illumination in the scene. Furthermore, with three or more illumination conditions, we can utilize photometric stereo for dense facial geometry reconstruction.

2 Previous Work

2.1 Passive capture

The simplest capture setup requires no specialized hardware and records the facial performance with a single video camera. In the absence of 3D cues, prior facial models such as 2D active appearance models [5] or 3D morphable models [3] can be used to constrain the recovered motion parameters. The quality of the recovered motion, however, is highly dependent on the training database. Generalized facial models trained on a large set of subject are capable of accurately categorizing emotions, but may miss fine details and motions unique to a specific subject.

Active appearance models were used on James Cameron's film "Avatar" to recover some eye motion from head-mounted camera data. Head-mounted cameras have also been used with the proprietary facial analysis software developed by the company Imagemetrics. Unfortunately, video from head-mounted camera is often characterized by sudden changes in illumination as the actor moves through the capture stage or rotates her head. In general, automated computer vision algorithms have difficulty distinguishing changes in facial expression from changes in illumination. Both rigs used by Imagemetrics and for "Avatar" are affected by moving ambient light despite using a visible LED as a fixed illumination source.

2.2 Stereo capture

Additional stereo cameras can be utilized to recover 3D geometry. A survey of stereo algorithms can be found in [19][20]. Commercially, a head mounted rig with four small high-definition cameras was developed by the company Imagemovers and first used on Robert Zemeckis's film "A Christmas Carol". For dynamic performances, stereo can be extended to multi-camera optical flow for tracking facial motion [4][1]. Beeler et al. [1] demonstrated a single-shot technique for high quality geometry using high resolution stereo cameras and a displacement map based on ambient shadowing of skin pores.

As with all passive techniques, stereo matching and optical flow rely on the natural texture of face, such as skin pores, to find corresponding points between photographs. While the face exhibits a wide range of geometric and texture detail at multiple scales, many of these features may not be visible under ambient illumination. In areas with insufficient texture, stereo and optical flow techniques rely on regularization which results in a loss of surface detail. Additional surface detail can be created by the application of skin makeup. Bickel et al. [2] applied colored makeup and used shape from shading to recover specific areas of wrinkling. MOVA's CONTOUR Reality Capture system uses fluorescent makeup and ultraviolet illumination to generate dense randomized facial texture. Applied makeup can be also seen as a form of motion capture marker.

2.3 Marker-based Capture

In the commercial world, marker-based motion capture remains by far the most popular solution for full-body and facial performance capture. Many different types of markers exist including passive retroreflective markers, coded LEDs, and accelerometers. As camera technology increases in speed and resolution, systems can identify denser data sets with more and smaller markers. While sparse points provide useful information about the large-scale shape of the face, they miss several critical regions such as fine-scale skin wrinkling, complex mouth contours, eye contours and eye gaze. Commercial productions require significant effort from animators to manually recreate missing motion detail. To remain faithful to the original performance, these artists rely on additional reference cameras, including the head-mounted cameras.

One of the first head-mounted cameras for facial performance capture was used on Robert Zemeckis's film "Beowulf". The camera was combined with electro-ocularography sensors that attempted to directly record nerve signals for eye muscles. Unfortunately the recorded signals were noisy and unreliable and could not be used without additional cleanup.

2.4 Structured-light capture

Active illumination approaches can recover geometric information without relying on natural features. Structured-light capture techniques correspond camera and projector pixels by projecting spatially varying light onto the face. Depth accuracy is limited by the resolution of the camera and projector. Different sets of illumination patterns have been optimized for processing time or accuracy. At one extreme, a single-shot noise pattern can be used with traditional stereo algorithms. An example of this is the popular Kinect controller for the Xbox game system which uses a hard-coded matching algorithm to achieve real-time depth but with very limited accuracy. Alternatively, a large set of sequential patterns could be used to fully encode projector pixel location. During a dynamic performance, there may be significant motion between subsequent illumination frames. Motion artifacts can be handled by either reducing the number of projected patterns [18][29] or explicitly shifting the matching window across time [28][6]. At this time, structured-light has not been used in head-mounted camera systems due to the size and weight of projection hardware.

2.5 Dynamic Photometric Stereo

Another form of active illumination is photometric stereo. Traditional photometric stereo [27] uses multiple point lights to recover surface orientation (normals) by solving simple linear equations. Unlike stereo and structured light techniques which recover absolute depth, surface orientation is equivalent to directly measuring the depth derivative. As a result, photometric stereo provides accurate local high-frequency information, but is prone to low-frequency errors [17]. Photometric stereo also has the advantage that it can be computed in real-time on standard graphics hardware [16]. As with structured light, it is desirable to reduce the total number of photographs. In his original paper [27], Woodham suggested a single shot approach where the different illumination directions are encoded in the red, green, and blue color channels. The drawback of this approach is that it assumes constant surface color. Recent papers have extended this idea using optical flow [7][12], white makeup [13], better calibration [11] or additional spectral color channels [10]. Ma et al. [14] formulated photometric stereo using four spherical gradients to minimize shadows and capture normals for the entire face. This has been used for dynamic facial performance capture [15][9] but requires a large lighting apparatus. In this paper, we show that photometric stereo is be used in a light-weight, portable device where lights and cameras are in close proximity to the subject.

3 Apparatus

The development of head-mounted capture systems has always been spurred by developments in tiny wearable cameras. Recently, several compact machine vision cameras have been introduced that support fast video capture. For our apparatus we use a Point Grey Grasshopper camera capable of VGA resolution video at 120fps while weighing only 100 grams. Upcoming cameras, such as as Point Grey Flea 3 and Basler Ace, will achieve high-definition video in an even smaller form-factor approximately the size of an ice cube. Of course,

high-speed cameras also generate more data which has to be streamed back to a computer. Our current system is tethered via Firewire 800 with a range of 4.5 meters. Untethered performances should be possible by using a battery pack and real-time data compression hardware to store data on the subject's body.

Figure 1: Head-mounted camera and LED light ring

Figure 2: (left) The LED ring, (center) a single-LED pattern, (right) a gradient LED pattern.

The target frame rate for the captured performance is 30fps, so our camera provides a budget of four captured images for each final output frame. We utilize these frames by time-multiplexing four distinct illumination conditions. Our light sources consist of 12 individually controlled Luxeon 5 LEDs that form a 5cm diameter ring encircling the lens (see Figure 2). As traditional photometric stereo assumes lambertian material properties, we mount crossed linear polarizers on the lights and lens to attenuate specular reflection from the face. The total LED ring weighs 2.7 grams. The camera and LED ring are then mounted to a helmet on the end of a 20cm arm (see Figure 1).

At 120Hz, flicker from the white LEDs is visible to the actor. Flicker could be greatly reduced by switching patterns at 360Hz while recording at 120fps with 1/360th sec exposure, capturing every third pattern. The resulting interspersed frames would match those captured at the slower switching speed, but may require additional light to compensate for the shorter exposure time. Distraction from the lights can be completely eliminated by using invisible near-infrared LEDs and leveraging the broad spectral sensitivity of CCD cameras. Infrared illumination provides reasonable results at the expense of some detail in the photometric normals due to increased subsurface scattering at longer light wavelengths. Infrared optics and polarizers are also more expensive than equivalent visible-spectrum components. A sample frame captured with infrared light can be seen in Figure 4

Figure 3: Four time-multiplexed illumination conditions using single visible LEDs.

Figure 4: Comparison of face lit by (left) visible and (right) near infrared gradients.

Different lighting configurations provide a tradeoff between surface normal accuracy and other artifacts such as shadowing. Using our ring, we test two different lighting setups: single LED point lights, and larger linear gradients that ramp across all twelve LEDs (Figure 2). Concentrated point lights have the advantage that they could be adapted to a smaller form factor. Alternatively, gradients using the full ring of lights provide more even illumination and reduce shadow artifacts.

When using both patterns, we follow the sequence of three illumination patterns with an unlit frame with all LEDs turned off. This frame is used to record and subtract any ambient light in the scene (Figure 3).

4 Motion Compensation

As the different lighting conditions are captured sequentially, there may be noticeable motion between adjacent video frames. It is difficult to track this motion directly between all four lighting conditions, as the changing illumination violates the brightness constancy assumption used in most optical flow algorithms. As in [23], we overcome this problem by computing flow between frames that share the same lighting condition. We compute flow between every fourth frame and and linearly interpolate to warp the remaining frames. At 120fps we found that the linear motion assumption does not create significant artifacts. The entire optical flow computation takes place in under a 1/30th of a second using GPU-based optical flow [24][25]. This is particularly useful when providing real-time feedback to the director or actor during the performance. Alternative CPU-based optical flow algorithms such as [22] could be used for off-line processing.

5 3D Reconstruction

As presented by Woodham [27], photometric stereo estimates surface orientation (normals) by analyzing how a surface reflects light incident from multiple directions. For lambertian reflectance, image intensity (I) can be expressed as the dot product of the lighting direction (L) and surface normal (N) scaled by the albedo (A) for each pixel in the image.

$$I = L \cdot NA \tag{1}$$

Given three observations of a pixel and the corresponding lighting directions, equation (1) can be solved by inverting the 3x3 matrix of known lighting directions. After multiplying the inverted matrix by the observed pixel values, the resulting vector's length is the surface albedo and the normalized vector is the estimated surface normal.

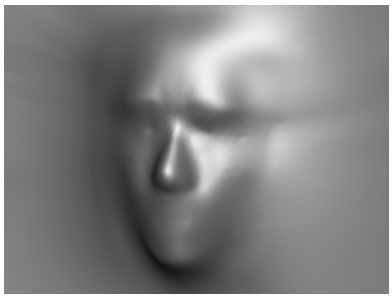

Figure 5: Smoothed template geometry used to initialize relative lighting directions and depth.

We physically measure the absolute LED positions relative to the helmet prior to each performance. Due to the near-proximity of the LEDs, lighting directions will vary for each pixel. We initialize these directions using a smoothed face model placed at the approximate position of the head within the helmet (Figure 5). During the performance, the per-pixel lighting directions and motion-compensated photographs are converted to surface normals using Equation (1). Our implementation uses a GPU fragment shader running in real-time.

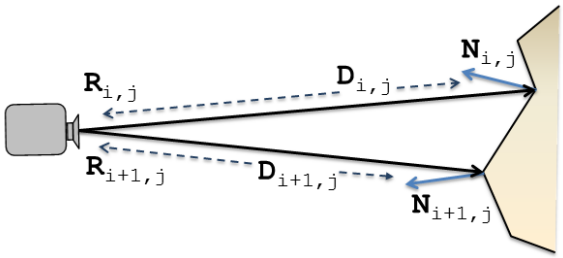

Figure 6: Updated gradients can be computed based on neighboring ray directions, estimated depth and surface normals.

Surface normals are also a measurement of the depth gradient. These gradients can be integrated across the face to recover the 3D geometry. Most normal integration methods assume an orthographic or distant camera, where the depth gradients (G_x, G_y) are given by the following equation:

$$\begin{aligned} G_x &= \frac{N_x}{N_z} \\ G_y &= \frac{N_y}{N_z} \end{aligned} \tag{2}$$

However, the head-mounted camera has a wide field of view so camera rays are not parallel. If we naively compute gradients using Equation (2), the integrated geometry will exhibit fisheye distortion where objects closer to the camera are too large relative to those further way. This effect can be reduced by calibrating camera intrinsics and computing gradients relative to the diverging camera rays. For a given pixel (i, j), the new depth gradients are a function of neighboring surface normals (N), ray directions (R) and the distance of each pixel from the camera (D).

$$\begin{aligned} G_x = D_{i+1,j}\left(1 - \frac{R_{i+1,j} \cdot N_{i+1,j}}{R_{i,j} \cdot N_{i+1,j}}\right) \\ -D_{i,j}\left(1 - \frac{R_{i,j} \cdot N_{i,j}}{R_{i,j} \cdot N_{i+1,j}}\right) \\ G_y = D_{i,j+1}\left(1 - \frac{R_{i,j+1} \cdot N_{i,j+1}}{R_{i,j} \cdot N_{i,j+1}}\right) \\ -D_{i,j}\left(1 - \frac{R_{i,j} \cdot N_{i,j}}{R_{i,j} \cdot N_{i,j+1}}\right) \end{aligned} \tag{3}$$

We reuse the smoothed template geometry to initialize the per-pixel depth (D). The corresponding integrated geometry will exhibit high-frequency detail from the surface normals and low-frequency shape from the template mesh. To generate more accurate geometry, we can update lighting directions and depth gradient estimates based on the integrated geometry, then iterate both the photometric stereo and normal integration stages.

5.1 Results

To test the system, we captured several performances using point light source and gradients and recovered surface normals, albedo texture, and integrated geometry. Figure 8 shows a samples sequence under point-light source illumination as the subject recites the line "The Five Wizards Jumped Quickly". Our system was able to capture the fast mouth motion associated with the different visemes as well as subtle eyes motion and nose twitches. The full performance can be found in the accompanying video.

In general, the reduced separation between gradient lighting patterns (shown in Figure 4) resulted in higher levels of noise in the surface normals (see video). As shadowed regions were relatively small, we found the point light sources produced the best results. In Figure 8 shadow artifacts can be seen as white albedo around the nostril and as a flattening of normals and geometry. These errors could be eliminated by explicitly detecting shadows and updating the integration constraints as in [8]. Our initial prototype was also unable to completely eliminate specular reflections in the infrared. These specular highlights, seen on the tip of the nose, produce incorrect spikes in the surface normals and geometry (see video).

The remaining low-frequency errors in the surface normals and geometry could be eliminated by combining photometric stereo with more conventional techniques. For example, a second camera could be mounted on the helmet for stereo matching or to triangulate motion capture markers. The resulting sparse geometry could then be used instead of the generic face template to initialize lighting directions and depth gradients.

The 3d shape information provided by our head-mounted camera opens multiple possibilities for driving a facial rig. Techniques designed to work with structured-light data or depth cameras such as [21] could be adapted to use depth from integrated surface normals. Alternatively, surface normals could be used directly as an additional channel of information in an active appearance model.

6 Future Work

We are continuing to refine our motion capture rig to further decrease weight and improve balance. One improvement is to mount the the camera directly to the helmet and use a convex mirror on the end of the arm (see Figure 7). The arrangement moves the center of mass towards the helmet and increases the effective distance of the camera. As with conventional head mounted cameras, this setup is susceptible to camera shake if the helmet is not locked securely to the head. Additional tracking markers could be placed on the arm and helmet to record this motion and stabilize the video relative to the face.

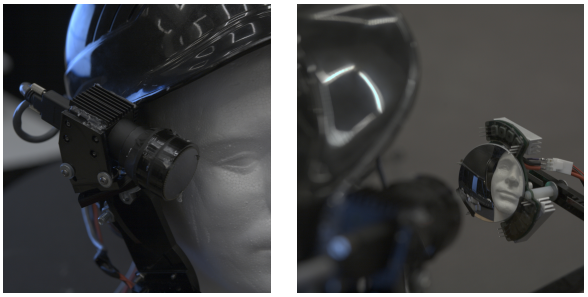

Figure 7: Alternate configuration using a camera and mirror to improve weight distribution on the helmet.

As video projectors are rapidly shrinking in size, power consumption, and cost, it may be possible to adapt other forms of active-illumination to a head-mounted form factor. The latest DLP-based Pico projectors from Texas Instruments are capable of frame rates above 120Hz and weighs 1.7gm. These projectors could serve as an alternate LED light source for photometric stereo, or generate structured-light patterns.

We are interested in conducting more extensive testing of our apparatus in real-world motion capture environments. Most motion capture systems are sensitive only to specific infrared bands, so there should be minimal interference between the head-mounted LEDs and the motion capture system. Photometric stereo is a natural addition to existing commercial head-cameras that already incorporate LED illumination. Given simultaneous high resolution facial data body markers, it should be possible to identify and study interesting correlations between facial and body motions.

6.1 Conclusion

As digital characters proliferate in films and interactive games, the line between what is real and what is artificial continues to blur. With existing motion capture technology, animators are required to reinterpret an actor's performance in order to accentuate key dynamics of a performance. This process is not only time-consuming, but also raises the artistic question of whose performance is being presented: the actor's or the animator's. Head-mounted photometric stereo captures an entire face, not just a sparse set of motion capture markers. While this information can already be used as a more accurate reference for animators, we hope that it will inspire new algorithms for automatically driving digital characters, while remaining faithful to the original performance.

References

[1] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics*, 29(4):40:1–40:9, July 2010.

[2] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy,

Figure 8: Results showing every 18th frame from a performance captured under single LED illumination. (row 1) Surface albedo recovered using photometric stereo, (row 2) surface normals with XYZ encoded as RGB, (row 3) integrated surface geometry.

H. Pfister, and M. Gross. Multi-scale capture of facial geometry and motion. *ACM Transactions on Graphics*, 26(3):33:1–33:10, July 2007.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH 99*, Computer Graphics Proceedings, Annual Conference Series, pages 187–194, Aug. 1999.

[4] G. Borshukov, D. Piponi, O. Larsen, J. P. Lewis, and C. Tempelaar-lietz. Universal capture: image-based facial animation for 'the matrix reloaded'. In *SIGGRAPH 2005 Courses*, 2005.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[6] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *2003 Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, pages 359–366, June 2003.

[7] B. De Decker, J. Kautz, T. Mertens, and P. Bekaert. Capturing multiple illumination conditions using time and color multiplexing. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2536 –2543, June 2009.

[8] C. H. Esteban, G. Vogiatzis, and R. Cipolla. Overcoming shadows in 3-source photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 2011.

[9] G. Fyffe, T. Hawkins, C. Watts, W.-C. Ma, and P. E. Debevec. Comprehensive facial performance capture. *Eurographics: Comput. Graph. Forum*, 30(2):425–434, 2011.

[10] G. Fyffe, X. Yu, and P. Debevec. Single-shot photometric stereo by spectral multiplexing. In *Computational Photography (ICCP), 2011 IEEE International Conference on*, pages 1 –6, April 2011.

[11] C. Hernandez and G. Vogiatzis. Self-calibrating a real-time monocular 3d facial capture system. In *Proceedings International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[12] H. Kim, B. Wilburn, and M. Ben-Ezra. Photometric stereo for dynamic surface orientations. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 59–72, 2010.

[13] M. Klaudiny, A. Hilton, and J. Edge. High-detail 3d capture of facial performance. In *3DPVT*, 2010.

[14] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Rendering Techniques 2007: 18th Eurographics Workshop on Rendering*, pages 183–194, June 2007.

[15] W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Transactions on Graphics*, 27(5):121:1–121:10, Dec. 2008.

[16] T. Malzbender, B. Wilburn, D. Gelb, and B. Ambrisco. Surface enhancement using real-time photometric stereo and reflectance transformation. In *Rendering Techniques 2006: 17th Eurographics Workshop on Rendering*, pages 245–250, June 2006.

[17] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Transactions on Graphics*, 24(3):536–543, Aug. 2005.

[18] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Transactions on Graphics*, 21(3):438–446, July 2002.

[19] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.

International Journal of Computer Vision, 47(1–3):7–42, Apr./June 2002.

[20] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pages 519–528, June 2006.

[21] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. *ACM Transactions on Graphics*, 30(4), July 2011.

[22] M. Weiss. Depth-discontinuity preserving optical flow using time-multiplexed illumination. Master's thesis, RWTH Aachen University, University of Southern California, 2007.

[23] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics*, 24(3):756–764, Aug. 2005.

[24] M. Werlberger, T. Pock, and H. Bischof. Motion estimation with non-local total variation regularization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010.

[25] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proceedings of the British Machine Vision Conference (BMVC)*, Sept. 2009.

[26] L. Williams. Performance-driven facial animation. In *Computer Graphics (Proceedings of SIGGRAPH 90)*, pages 235–242, Aug. 1990.

[27] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[28] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Transactions on Graphics*, 23(3):548–558, Aug. 2004.

[29] S. Zhang and P. Huang. High-resolution, real-time three-dimensional shape measurement. *Optical Engineering*, 45(12), 2006.