Video Normals from Colored Lights

Gabriel J. Brostow, Carlos Hernández, George Vogiatzis, Björn Stenger,
and Roberto Cipolla, *Members, IEEE*

Abstract—We present an algorithm and the associated single-view capture methodology to acquire the detailed 3D shape, bends, and wrinkles of deforming surfaces. Moving 3D data has been difficult to obtain by methods that rely on known surface features, structured light, or silhouettes. Multispectral photometric stereo is an attractive alternative because it can recover a dense normal field from an un-textured surface. We show how to capture such data, which in turn allows us to demonstrate the strengths and limitations of our simple frame-to-frame registration over time.

Experiments were performed on monocular video sequences of un-textured cloth, and faces with and *without* white makeup. Subjects were filmed under spatially separated red, green, and blue lights. Our first finding is that the color photometric stereo setup is able to produce smoothly varying per-frame reconstructions with high detail. Second, when these 3D reconstructions are augmented with 2D tracking results, one can both register the surfaces and relax the homogenous-color restriction of the single-hue subject. Quantitative and qualitative experiments explore both the practicality and limitations of this simple multispectral capture system.

Index Terms—Photometric stereo, multispectral, single view, Video Normals.

1 Introduction

The modeling of dynamic cloth geometry is increasingly based on computer vision techniques [1], [2], [3], [4], [5]. Both cloth and faces entail complex underlying dynamics that motivate capturing motion data from the real world whenever possible.

Existing algorithms one might employ for capturing detailed 3D models of moving cloth or skin include multiple view stereo [6], photometric stereo [7], [8], and laser based methods [9]. However, most of these techniques require that the subject stand still during the acquisition process, or move slowly [10]. Another substantial challenge is that even starting from a sequence of 3D scans of the deforming object, registration is necessary to produce a single 3D model, suitable for CG animation or further data analysis, such as used in [11] and [12].

The technique proposed here for acquiring complex motion data from real moving cloth and faces uses a highly practical setup that consists of an ordinary video camera and three colored light sources (see Figure 1). The key observation is that in an environment where red, green, and blue light is emitted from different directions, a Lambertian surface will reflect each of those colors simultaneously without any mixing of the frequencies. The quantities of red, green, and blue light reflected are a linear function of the surface normal direction. A color camera can measure these quantities, from which an estimate of the surface normal direction can be obtained. By applying this technique to a video sequence of a deforming object, one can obtain a sequence of normal maps for that object which, in turn, allows us to make the following contributions:

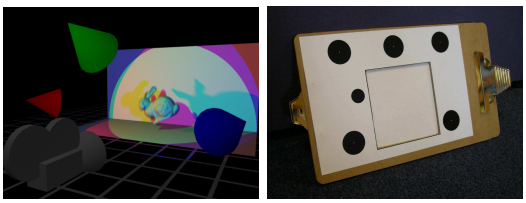


Fig. 1: **Setup and calibration board.** Left: a schematic representation of our multispectral setup. Right: Attaching two boards with a printed calibration pattern results in a planar trackable target for computing the orientation of the pattern's plane. The association between color and orientation can be obtained from a cloth sample inserted in the square hole between the boards.

1) A simple acquisition setup for acquiring high-detail, per-frame reconstructions.
2) A simple calibration procedure that extends this technique to human faces.
3) A simple registration approach for real deforming surfaces with folds and creases, based on optical-flow.
4) An algorithm for detecting self-shadows.
5) An application of our method for 'dressing' a virtual character with real moving cloth.

Here, we apply our newest work for relaxing the need for gray materials [13] to extend our previous work [14] with (i) a new self-shadow detection algorithm, (ii) experiments on a rigid object for quantitative comparisons, and (iii) qualitative experiments to show the problems with registration and of using non-Lambertian surfaces. Video and calibration data

- *G. Brostow is with University College London.*
- *C. Hernández is with Google Inc.*
- *G. Vogiatzis is with Aston University.*
- *B. Stenger is with Toshiba Research Europe Ltd.*
- *R. Cipolla is with the University of Cambridge.*

from our experiments is provided online[1].

2 Related work

The animation and capture of cloth and face deformations is approached from various perspectives, and we review the most relevant ones with regard to the proposed technique.

2.0.0.1 Texture Cues: White and Forsyth [4], [5] and Scholz *et al*. [3] have presented work on using texture cues to perform the specific task of cloth capture. Their methods are based on printing a special pattern on a piece of cloth and capturing video sequences of that cloth in motion, usually with multiple cameras. The estimation of the cloth geometry is based on the observed deformations of the known pattern as well as texture cues extracted from the video sequence. The techniques produce results of good quality but are ultimately limited by the requirement of printing a special pattern on the cloth which may not be practical for a variety of situations. In the present work, we avoid this requirement while producing detailed results.

Pilet *et al*. [1] and Salzmann *et al*. [2] proposed a slightly more flexible approach where one uses the pattern already printed in a piece of cloth, by presenting it to the system in a flattened state. [15] were among the first innovators of such approaches. Using sparse feature matching, the pattern can be detected in each frame of a video sequence. Due to the fact that detection occurs separately in each frame, the method is quite robust to occlusions. However, the presented results dealt only with minor non-rigid deformations.

2.0.0.2 Photometric Stereo: Photometric stereo [16] is one of the most successful techniques for surface reconstruction from images. It works by observing how changing illumination alters the image intensity of points throughout the object surface. These changes reveal the local surface orientations. This field of local surface orientations can then be integrated into a 3D shape. State of the art photometric-stereo allows uncalibrated light estimation [8], [17] and can cope with unknown albedos [18], [19]. The main difficulty with applying photometric stereo to deforming objects lies in the requirement of changing the light source direction for each captured frame, while the object remains still. This is quite impractical when reconstructing the 3D geometry of a *moving* object, though Ma *et al*. [20] have built an impressive dome that uses structured and polarized multiplexed lighting to capture human faces. Still constrained by multiplexing, Vlasic *et al*. [21] demonstrated a multi-view system with eight 240Hz cameras and 1200 individually controllable light sources to capture geometry similar to our own. We show how multispectral lighting allows one to essentially capture three images (each with a different light direction) in a single snapshot, thus making per-frame photometric reconstruction possible and very accessible.

To really explore the limitations of our system, we also capture highly deforming human faces. The newest works by Ma *et al*. [22] and Wilson *et al*. [23] are among the highest quality face capture systems, in part because they build precise stages to capture both photometric stereo and precise depth. [22] is close to the ideal situation in all three ways, where photometric stereo captures detailed normals, projected structured light patterns capture accurate depth, and feature-tracking with extra cameras provides excellent landmarks for registration over time. They show how marker-based tracking can yield almost as high a quality facial animation, thanks to training a model in the heavily instrumented studio. Since heavy multiplexing was keeping them at a maximum of 30fps, [23] used high quality stereo cameras without the structured light to compute good depths, and added a new flow-based tracking to compensate for interframe motion. It could be interesting to extend our approach to use high quality stereo cameras in the future.

2.0.0.3 Colored and Structured Lights: The earliest related works are also the most relevant. The first reference to multispectral light for photometric stereo dates back 20 years to the work of Petrov [24]. Ten years later, Kontsevich *et al*. [25] actually demonstrated an algorithm for calibrating unknown color light sources and at the same time computing the surface normals of an object in the scene. They verified the theory on synthetic data and an image of a real egg. Drew and Kontsevich [26] even present evidence suggesting that the famous Lena photo was made under spectrally varying illumination. Woodham [27] also demonstrated that multi-spectral lighting could be exploited to obtain at least the normals from one color exposure. Also similar to our approach, his normals could be computed robustly when some self-shadowing was detected. Without using a calibration sphere made of the same material as the subject, we take a practical approach for calibration, and the same orientation-from-color cue, to eventually convert video of un-textured cloth or skin into a single dense surface with complex changing deformations. For the simplified case of a rigid object, [28] is using this principle to capture relief details by pressing it against an elastomer with a known-albedo skin.

The parameters needed to simulate realistic cloth dynamics were estimated from video by projecting explicitly structured horizontal light stripes onto material samples under static and dynamic conditions [29]. This system measured the edges and silhouette mismatches present in real vs. simulated sequences. Many researchers have utilized structured lighting, and Gu *et al*. [30] even used color, although their method is

1. http://mi.eng.cam.ac.uk/research/projects/VideoNormals/

mostly for storing and manipulating acquired surface models of shading and geometry. Weise *et al*. [31] leads the structured light approach, and has some advantages in terms of absolute 3D depth, but at the expense of both spatial and temporal sampling, e.g. 17 Hz compared to our 60 Hz (or faster, limited only by the camera used). Zhang *et al*. [32] also presented a complete system that uses structured light for face reconstruction.

2.0.0.4 Multi-View Registration with 3D Templates: Sand *et al*. dispensed with special lighting but leverage markered motion capture and automatic silhouettes to deform a human skeleton and body template [33]. The numerous and recent progress in cloth animation is based on this concept of matching a specially-built 3D template mesh to videos filmed in elaborate multi-camera systems with studio lighting (or structured lighting as in [34]). Bradley *et al*. [35] opt for a simple manual step for template-creation, that then hinges on the video resolution to create wrinkles. De Aguiar *et al*. [11] use a single 360° laser-scan to create a very precise template, and then address the challenge of preserving those wrinkles and folds while the actor moves around. Vlasic *et al*. [12] have a very similar process, that also starts with a laser scan or with a template made by Starck and Hilton [36]. Our technique, on the other hand, expects no prior models of the cloth being reconstructed. Instead, our algorithm could eventually be extended to be a precursor stage for those systems. There are potentially benefits if they used time-varying templates with our level of detail, instead of static ones.

2.0.0.5 Registration With and Without Articulation: Registration is *not* the emphasis of our research, but it is an inherent part of using our time varying surfaces in applications. Works in this area focus on the registration problem itself, except [37] who couple registration with their own capture system. Unlike ours, their approach both requires and benefits from i) a pre-made smooth template of the body, ii) an articulated skeleton of each subject which is used in their standard articulated-motion-capture framework, and iii) a multi-camera studio. Like most registration techniques, including our own, any assumptions about smoothly changing normals can ruin the high quality normal fields that may have been recovered. This technique winds up smoothing and interpolating normals over a window of five frames, precluding capture of normals for examples with flapping cloth, like our pirateShirt sequence visible online.

The focus of [38] is on articulated or piecewise-rigid shapes, where there is a known number of limbs, and they are pre-segmented for at least one depth-image. For this technique to succeed, consecutive frames must be close enough to give classic ICP a good initialization, which can be viewed as similar to our assumption about local flow on Video Normals. Other registration techniques for articulated shapes are fully automatic, such as [39] who discretize pose space and then seek out favorite transformations that align large sections of the two point clouds. We found the spin-images descriptor [40] to be brittle for single-view surface scans, but [41] is able to make skeletons out of similar data, enabling [42] to demonstrate good registration on synthetic and man-made shapes.

Multiple techniques now attempt to register the available point clouds (or volumetric scans [43]) in batch mode instead of online. Mitra *et al*. [44] successfully registers many scans of stiff objects all at once, instead of using a sequence of ICP-steps chained together. Their extensions for deformable bodies assume very limited degrees of freedom, which is not the case with our data, and they revert to optimizing just one time slice, unlike the main 4D funcion. They emphasized how errors crop up for them because of incorrect normals and non-rigid motion, which are exactly the problems we are addressing.

Also in the family of batch registration algorithms, Süßmuth *et al*. [45] and Wand *et al*. [46] have shown very nice general-purpose approaches that make few assumptions, and are mostly just limited by memory capacity. They have even registered sequences of faces as long as 150 frames. This is particularly hard with just points that are not parameterized in a graph with edges. [45] embeds the series of 3D point clouds in a 4D implicit function, and apply an EM-type optimization to find mesh deformations that prefer rotation and keep close to the positions of the point-clouds in the immediate temporal neighborhood. Their algorithm can be seen as parallel to the registration steps of our own, and possibly more extendible, in that their embedding of the point clouds in an implicit function (though costly) could theoretically be extended to allow the extracted meshes to change topology over time. [46] presents an impressive optimization system for computing a single shape and its time-varying deformation function from a sequence of point clouds (as many as 201 frames). The point clouds must overlap substantially to allow registration of temporal neighbors, but holes and gaps can come and go, and the technique eventually merges the deforming scans into a single urshape, with better coverage than individual scans. At the heart of the algorithm is a meshless volumetric deformation model with an energy function that allows consistent parts of multiple point clouds to be aligned with each other. Hierarchical processing in the time domain leads to a globally consistent solution, which is attractive compared to our frame-to-frame registration, except for the memory constraints and running times. We have our own data acquisition process that rivals what the authors of this paper assume as input, and we explicitly detect shadows and apply no data-culling. Our registration does accumulate error, but has a simpler regularization that does not penalize volumetric, velocity, or acceleration changes. So speaking quite

broadly, ours is "fast and cheap", while theirs is slow but good for many of the same situations we care about. Qualitative evaluation of the resulting videos is necessary to assess the amount of detail retained in our respective registered models.

3 Depth-map video

In this section, we follow the notation of Kontsevic *et al*. [25]. For simplicity, we first focus on the case of a single distant light source with direction $\mathbf{l} = [\, l_1 \; l_2 \; l_3 \,]^{\mathrm{T}}$ illuminating a Lambertian surface point $\mathbf{x}$ with surface normal $\mathbf{n}$. Let $S(\lambda)$ be the energy distribution of that light-source as a function of wavelength λ and let $\rho(\lambda)$ be the spectral reflectance function representing the reflectance properties at that surface point. We assume our camera consist of multiple sensors (typically CCD's), sensitive to different parts of the spectrum. If $\nu_i(\lambda)$ is the spectral sensitivity of the i-th sensor for the pixel that receives light from $\mathbf{x}$, then intensity measured at that lone sensor is $r_i = \mathbf{l}^{\mathrm{T}}\mathbf{n} \int S(\lambda)\rho(\lambda)\nu_i(\lambda)d\lambda$, or in matrix form

$$\mathbf{r} = M\mathbf{n}, \tag{1}$$

where the (i, j)-th element of the 3-column M is

$$m_{ij} = l_j \int S(\lambda)\rho(\lambda)\nu_i(\lambda)d\lambda. \tag{2}$$

To solve for the 3 unknowns of $\mathbf{n}$, M must be rank 3, meaning 3 or more rows of r_i (*i.e.* sensors) are required. Actually, even with 3 sensors, M would be of rank 1 when using just one light source, because the per-sensor dot products are not linearly independent. When more light sources are added, if the system is linear and $\mathbf{l}^{\mathrm{T}}\mathbf{n} \geq 0$ still holds for each light, the response of each sensor is just a sum of the responses for each light source individually, so we retain (1) but with

$$M = \sum_k M^k, \tag{3}$$

where M^k describes the k-th light source. Therefore, in the absence of self occlusions, three sensors and a minimum of three different lights need to be present in the scene for a pixel's M to be invertible. If the surface is uniformly colored (constant albedo), then the reflectance $\rho(\lambda)$ and consequently M will be fixed across all un-occluded locations.

Equation (1) establishes a one-to-one mapping between an RGB pixel measurement from a color camera and the surface orientation at the point projecting to that pixel. Our strategy uses the inverse of this mapping to convert a video of a deformable surface into a sequence of normal maps.

3.1 Setup and calibration

Our setup consists of a color video camera and three light sources which have been filtered with red, green and blue filters respectively. The camera is placed 2.5-5m away from the target object. The light sources are at a similar distance, not colinear, aimed at the target, and separated by about 30 degrees from one another. The filming occurs in a dark room with minimal ambient light. Figure 1 (left) describes this schematically.

In [25] and [47], methods were proposed for the estimation of the linear mapping M of equation (1) from the image itself, using the constraints of uniform albedo and surface integrability that must be satisfied by the normal map. However the results obtained with these techniques can be unsatisfactory, especially in situations where the target object does not have a wide range of surface orientations (e.g. if it is mostly planar). We prefer to estimate the mapping by employing an easy-to-use calibration tool (Figure 1, right) similar to the one used in [48]. The pattern is planar with special markings that allow the plane orientation to be estimated. By placing the cloth in the center of the pattern, we can measure the color it reflects at its current orientation. We thus obtain a set of $(\mathbf{r}, \mathbf{n})$ pairs from which the mapping M is estimated using linear least squares [14].

3.2 Depth from Normals

By estimating and inverting the linear mapping M linking RGB values to surface normals, we can convert a video sequence captured under colored light into a *video of normal-maps*. Each normal map is integrated independently for each frame using a Fast Fourier Transform (FFT) method [49]. At the end of the integration process, we obtain a *video of depth-maps*.

4 Human Face Normals

The motion of cloth can be dynamic and intricate, but cloth is also flexible and easily used in our original flat-surface calibration method [14]. Here we extend the previous approach to reconstruct moving human faces.

A trivial extension for capturing Video Normals of moving faces is to fully apply makeup to the skin, and then use the same makeup on a flat surface in the calibration board of Figure 1. Such a calibration makes the assumption that the makeup is matte and evenly applied. While approximate and slightly inconvenient for the actor, this simple approach is surprisingly effective (see Figure 2).

It is worth noting that some existing facial scanning [50] and motion capture systems can already produce excellent results, but often at the cost of having a more complicated setup. Ma et al. [20] use polarized spherical gradient illumination patterns and

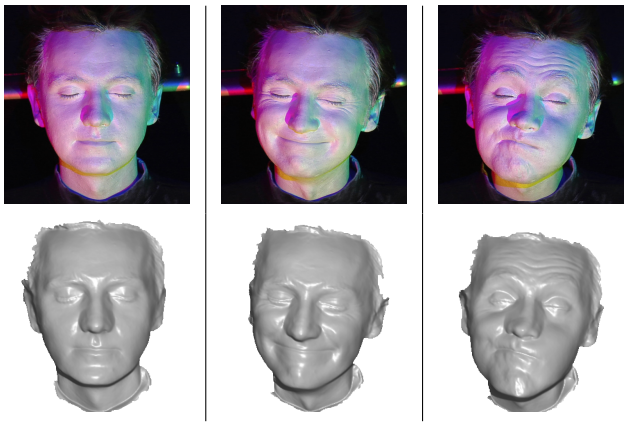


Fig. 2: **Applying the original algorithm to a face with white makeup.** Top: example input frames from video of an actor smiling and grimacing. Bottom: the resulting integrated surfaces.

multiplexing to recover detailed surface geometry. Furukawa and Ponce [51] have recently introduced a new tangential rigidity constraint for registration, but also rely on multiple synchronized cameras. Bradley et al. [52] recently showed excellent results with a 14-camera system with special lighting that allowed them to register geometry and textures using a stereo flow-based technique, similar to the one we use here for single-view capture. While they succeed by tracking highly detailed texture, we are able to track the video of normals, though we take no face-specific steps to counteract drift, which eventually leads to our accumulation of errors.

Good facial expression capture should not depend on makeup. The calibration step is extended, on the basis of [13], to cope with unpainted faces, and more generally, with single-hue objects that can be rotated in front of the camera without significant deformation. In practice, during this calibration step, the makeup-free actor need only hold some expression while turning their head all the way to the left and right. The head itself is used as a rigid calibration object, and the per-frame pose and 3D shape are estimated in order to obtain M, the skin's response to this arrangement of multi-colored illumination.

The first step is to establish the changing pose of the head. Although skin can appear mostly smooth, the blue channel of facial skin shows fairly distinct (though sparse) trackable features. The 3D pose of these points on a rigid object is computed from the 2D tracks using established SfM algorithms [53]. We feed our own 2D tracks to the Boujou [54] software, producing the relative pose between the camera and each frame of the head. If 2D tracks are not available, silhouette-based calibration methods such as [55] or [56] can serve this purpose.

The second step uses the poses to help estimate the shape of the head, to an extent slightly better than a visual hull. We apply the silhouette and stereo fusion technique of [57] because it is simple and reliable. Reasonable alternatives exist for this stage, including [58] and [59]. The expectation here is only that the surface patches with a given world-orientation have a similar color overall, so the recovered head model's shape can be approximate. This initial head geometry is shown in Figure 3(B).

In the third step, the head's poses and approximate geometry are used to compute the illumination directions and intensities. Here, instead of the previous calibration of the 3×3 M matrix using a flat material sample, we use the estimated head model itself. Unlike Lim *et al.*'s reconstruction algorithm [17], we do not assume that all projected 3D surfaces are equally informative of illumination. We follow the RANSAC-based formulation of [8], where lighting is estimated from partially correct geometry. Our algorithm randomly selects a fixed number of points on the surface and uses their corresponding pixel intensities to hypothesize an illumination candidate. All surface points are then used for testing this hypothesis. This process is iterated and the candidate with the largest support is selected as the illumination estimate. This is more robust to both inaccurate geometry and inconsistent hue, because an illumination hypothesized based on an unfortunate choice of three points on the head mesh will receive fewer votes and appear as an unusual outlier compared to choices from the dominant color. For a pure Lambertian surface and distant point light source model, only three points are required to estimate illumination. However, the approach can easily cope with more complex lighting models. For example, a first order spherical harmonic model (3×4 matrix) could be estimated from four points. This approximation is equivalent to a distant point light source with ambient lighting. Figure 3 shows sample input and output frames from a longer face sequence without the use of the calibration board or any face makeup.

5 Tracking the Surface

While the *video of depth-maps* representation can be adequate for some applications, for texture mapping, points on different depth maps must be brought into correspondence. Figure 10 (second row) shows the failure of directly texture-mapping each depth-map of moving cloth without any registration. As mentioned in Section 2, one could choose to register the time-varying surfaces using one of many available algorithms, based on articulations, speed, or subject-specific constraints. Instead, we showcase the spatio-temporal detail of the points derived from Video Normals by doing simple frame-to-frame registration that is not limited by memory constraints when processing long sequences. We use optical flow, precisely because it relies on good texture details, and advect the first point cloud in experiments using two different registration optimizations. Let $z^t(u, v)$ denote the depth-

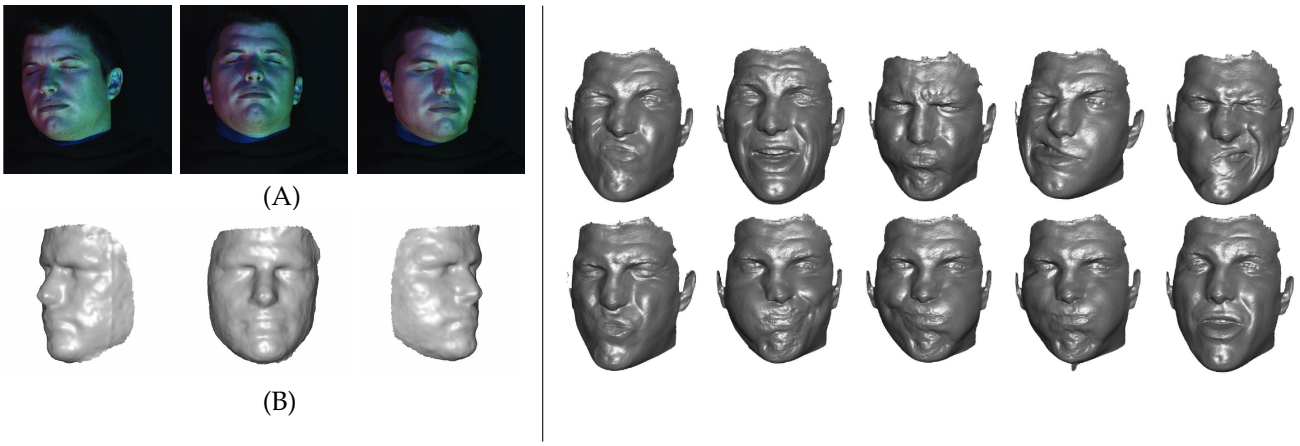


Fig. 3: **Face sequence without makeup.** Our calibration technique builds on multi-view reconstruction and lighting estimation (see Section 4). It is made possible by first moving the head around with a fixed expression (A). The initial recovered head geometry, shown in (B), is only approximate. The integrated surfaces are shown on the right using the self-shadow processing method of [60].

map at frame t. Our deformable template is the depth-map at frame 0, and is a dense triangular mesh with edges $\mathcal{E}$ and vertices $\mathbf{X} = \{\mathbf{x}_i^0\}$,

$$\mathbf{x}_i^0 = \left(u_i^0, v_i^0, z^0\left(u_i^0, v_i^0\right)\right), \quad i = 1 \ldots N. \tag{4}$$

Similarly to [61], the deformations of the template are guided by the following two competing constraints:

- the deformations should be compatible with the frame-to-frame 2D optical flow of the original video sequence,
- the deformations should be locally as rigid as possible.

5.1 2D Optical flow

We begin by computing frame-to-frame optical flow in the video of normal-maps. A standard optical flow algorithm is used for this computation [62], which for every pixel location (u, v) in frame t, predicts the displacement $\mathbf{d}^t(u, v)$ of that pixel in frame $t + 1$. Let (u^t, v^t) denote the position in frame t of a pixel which in frame 0 was at (u^0, v^0). We can *advect* $\mathbf{d}^t(u, v)$ to estimate (u^t, v^t) using the following equation from [33]:

$$(u^j, v^j) = (u^{j-1}, v^{j-1}) + \mathbf{d}^{j-1}\left(u^{j-1}, v^{j-1}\right), \quad j = 1 \ldots t. \tag{5}$$

If there were no error in the flow and our template from frame 0 had perfectly deformed to match frame t, then vertex $\mathbf{x}_i^0$ of the template would be displaced to point

$$\mathbf{y}_i^t = \left(u_i^t, v_i^t, z^t\left(u_i^t, v_i^t\right)\right). \tag{6}$$

5.2 Regularization

Simply moving each template vertex to the 3D position predicted by optical flow can cause stretching and other geometric artifacts like the ones displayed in Figure 10 (third row). This is due to accumulated error in the optical flow caused in part by occlusions. We tried two different regularization techniques. The first, described in more detail in our original paper [14], requires that translations applied to nearby vertices are as similar as possible. This is achieved by finding the $\hat{\mathbf{y}}_i$'s that optimize the energy term $E = \alpha E_D + (1-\alpha) E_R$. Here, α determines the degree of rigidity of the mesh, E_D is the data term, and E_R measures the dissimilarity of translations being applied to neighboring vertices. Reasonably good registration results are shown at the bottom of Figure 10.

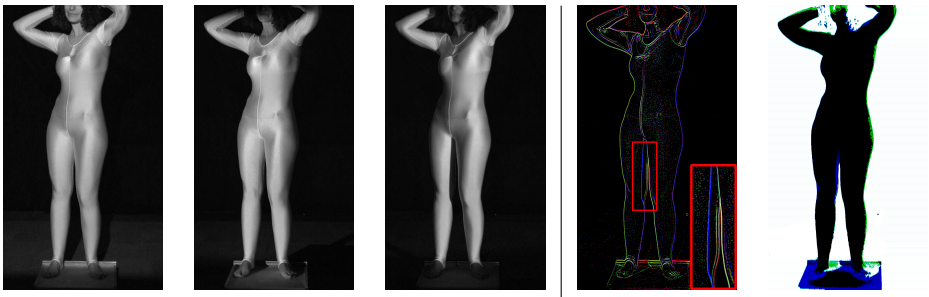
The alternative regularization technique is similar to the alignment-by-deformation of Ahmed *et al*. [63], and is based on Laplacian coordinates [64]. Unlike [63], we use the computed flow instead of SIFT features with adaptive refinement. Given the fine grid connection graph of $\mathbf{X}$, we make the $N \times N$ mesh Laplace operator $\mathbf{L}$, and apply it to the points from the template to convert them to Laplacian coordinates, $\mathbf{Q} = \mathbf{L}\mathbf{X}$. $\mathbf{Q}$ now encodes the high spatial frequency details of $\mathbf{X}$ and ignores its absolute coordinates. $\hat{\mathbf{Y}}$, the least-squares optimal absolute coordinates in the next frame, is computed by solving the linear equation

$$\begin{pmatrix} \mathbf{L} \\ \beta \mathbf{I}_N \end{pmatrix} \hat{\mathbf{Y}} = \begin{pmatrix} \mathbf{Q} \\ \beta \mathbf{Y} \end{pmatrix}, \tag{7}$$

which trades off the Laplacian coordinates against the results of tracking, using a similar rigidity parameter β. Section 7 describes the qualitative evaluation of how long each of the two regularization approaches tracks our Video Normals through large deformations before eventually falling off. In all the experiments, α was set to 0.9 and β to $1e-3$.

6 Self-Shadowing

So far, the algorithm computes Video Normals at each pixel in a given frame, independently of its neighbors in that frame. Unfortunately, it is inevitable that another part of the subject can come between the light and the camera, causing a self-shadow. This is



(A) (B) (C) (D) (E)

Fig. 4: **Spandex self-shadow images.** (A-C) are the red, green, and blue components of the recorded frame, while (D) shows the edges detected by the Laplacian filter. Note the prominent blue line running down the right leg, where the blue light cast a shadow. (E) shows where each of the lights cast its color shadow, except that the background has already been turned off.

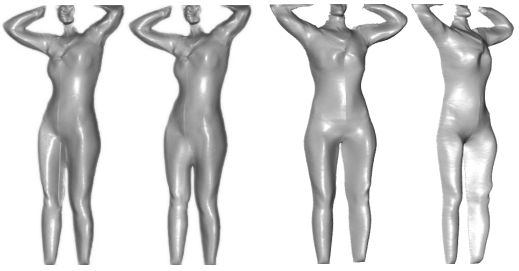
also a problem for regular photometric stereo, though there are potentially fewer self-shadows induced by one light source than by three. The three distributed lights however, offer a new opportunity that can be exploited to partly compensate when computing normals for shadowed surface patches.

For the first time in the algorithm, we consider the spatial relationship of the pixels in an image. When a photograph is considered as a composite of reflectance and illumination, Sinha and Adelson [65] observed that illumination varies more smoothly and is less likely to align with reflectance changes. Though we must contend with three sources of illumination, the three-channel video camera allows us to examine each light in turn, while reflectance changes were constrained from the outset. This justifies the use of a simple Laplacian edge-detector in each of the color channels of captured frame F_{RGB}. The resulting per-channel edges are pictured, with increased contrast for illustration, in Figure 4D.

Per-channel edge pixels are analyzed in turn to determine gradient orientation. We compute and quantize orientation by checking along each of the eight cardinal directions, at a distance of ± 2 pixels. Pixels whose gradient magnitude falls below a threshold τ are rejected. Adjoining pixels whose direction agrees are grouped into connected components, and we found empirically that for our footage, components with fewer than 20 pixels could safely be rejected at the conservative setting of $\tau = 5\%$. These parameters could change for filming under different conditions, to match the overall brightness of the average F.

The remaining gradient pixels are used as seeds for a conservative flood-filling algorithm which expands to neighbors whose intensity is equal or darker. With shadowed-pixels in each channel of F labeled, we compute a lookup visibility mask for each pixel, indicating which channels are present, if any. A dark backdrop was enough to insure that our algorithm labeled not only the correct regions on the actors as having two, one, or no discernable self-shadows, but also the surrounding scene as having all three shadows.

Finally, the parts of a surface that are self-shadowed by just one light source (*i.e.* $k = 2$) can now be processed specially to compensate for the missing channel of information (see Figure 5(A-B)). Onn and Bruckstein [66] addressed precisely this situation when dealing with two-image photometric stereo. The same ambiguity exists whether two gray-scale images are available, or when given F_{RGB} of a surface illuminated by just two colored lights. The local surface is constrained to have one of two possible orientations, corresponding to the two acceptable roots of a quadratic equation. Having classified the pixels as shadowed from a particular light, we choose the root whose normal is locally continuous with the unshadowed surface, under a constant albedo assumption. Figure 5(A-B) illustrates the effect of this improvement on the integrated surface. For the less obvious improvement for dealing with self-shadows (once found) and complicated albedo, see [60].



(A) (B) (C) (D)

Fig. 5: **Self-shadowing & the Lambertian assumption.** (A-B): Integrating the surface normals where all pixels are treated equally vs. using our self-shadow detection and correction (Section 6). The difference is most pronounced above the model's left knee. Separate from the matter of self-shadowing, (C-D) show a limitation of our system. Since the cloth violates our Lambertian assumption, the integrated surface of a different pose looks convincing from the front (C), but not from the side (D).

7 Experiments

Our experiments use real-world subjects filmed using a color video camera with resolution of either 1280×720 or 1024×1024 at 60fps. Since reconstruction consists of a matrix-vector multiplication followed by

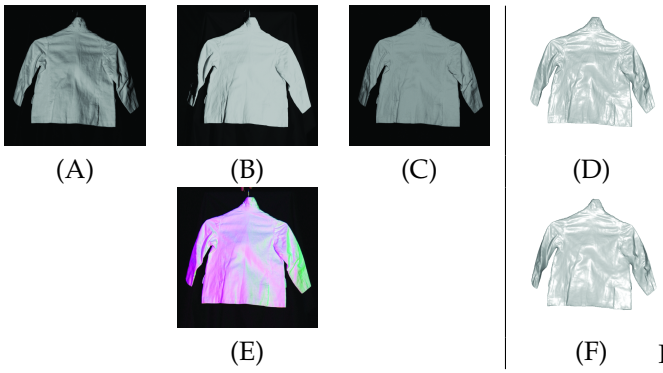


Fig. 6: **Comparison with photometric stereo.** (A-C) show three grayscale images captured by a digital camera, each taken under a different illumination, providing the input to a classic photometric stereo reconstruction [16] shown in (D). (E) shows a frame from a jacket sequence, where the same object is illuminated *simultaneously* by three different colored lights. Our algorithm only uses one such frame to generate the surface mesh shown in (F). Note that both algorithms give very similar results, but only the new one (bottom row) can work with video since only one frame is required to obtain a reconstruction. As a quantitative comparison, the average error between both reconstructions is only **1.4%** of the bounding box diagonal.

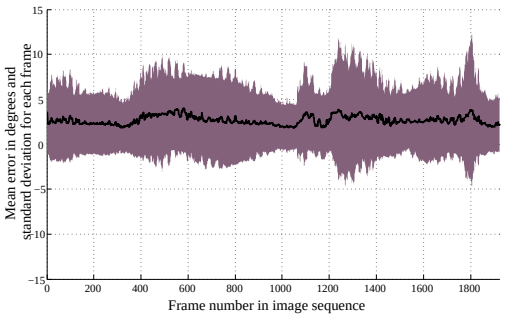


Fig. 7: **Cylinder reconstruction evaluation.** A rigid cylinder was moved in front of the camera and the geometry estimation was evaluated quantitatively. A best-fit cylinder geometry is computed for the sequence, so that for the cylinder's pose in each frame, we know the ideal normal-field. The plot shows the per-frame mean and standard deviation of the distance between the ideal and the estimated normals in degrees, as a function of time. The overall mean error was $2.67°$.

a Poisson integration [67], our FFT-based integration implemented with CUDA libraries produces depth-maps at 60 Hz. Computation times were on the order of 8 additional seconds for each registration of the mesh to the current frame. If the shadow correction algorithm from [60] is used, then the Poisson integration is about 10 seconds per frame. The sweater sequence meshes are 365k triangles and 183k vertices, while the makeup-free face mesh is 611k triangles and 307k vertices. Computations were carried out on a 2.8Ghz Pentium 4 processor with 4Gb of RAM and an nVidia GeForce 8800.

7.1 Quantitative comparisons

To evaluate the accuracy of the per-frame depth-map estimation, we first reconstructed a static object (a jacket) using classic photometric stereo with three images each taken under different illumination. The same object was reconstructed using a single image, captured under simultaneous illumination by three colored lights, using our technique. Figure 6 shows the two reconstructions side by side. The results look very similar and the average distance between the two meshes is only 1.4% of the bounding box diagonal. This demonstrates that equation (1) works well in practice. It is worth noting that even though photometric stereo achieves comparable accuracy, it cannot be used on a non-static object whose shape will change while the three different images are captured.

We have a further measure to quantitatively evaluate our technique. A rigid cylindrical object was wrapped in smooth paper, and moved in front of the camera for 30 seconds, exploring all six degrees of freedom. A best-fit cylinder geometry is computed for the sequence, so that for the cylinder's pose in each frame, we know the ideal normal-field, against which the Video Normals field is measured. In Figure 7, each frame's mean normal-vector error in degrees and standard deviation are plotted. Overall, the mean error was $2.67°$, and the standard deviation was 4.29. Our test-frames, code for evaluating them, and per-frame scores are online, with the aim of encouraging more meaningful algorithm comparisons, when possible.

7.2 Qualitative tests of cloth and face

For the third experiment shown here, a model wearing a white sweater was filmed dancing under our multispectral illumination setup (see first row of Figure 10). For qualitative purposes, in Figure 9 we show several views of frame #380 without the texture map and in high resolution (the mesh consists of 180k vertices). The images clearly show the high frequency detail of the sweater. To the best of our knowledge, this is the only method able to reconstruct deforming cloth with such detail. However, as expected, materials that are far from Lambertian exhibit noticeable artifacts, as in Figure 5(C-D).

We used this sequence to evaluate the original mesh regularization algorithm of Section 5 by texture mapping the deforming sweater. Figure 10 shows several approaches to mesh registration starting with no registration at all (second row), registration using the advected optical flow alone (third row) and the effect of regularizing optical flow with the rigidity constraint (fourth row). This last approach is seen to outperform the others as it manages to track the surface for more than 500 frames.

The fourth experiment explores tracking the much more challenging deforming face sequence from Figure 3. Rows of Figure 8 show different frames from among a sequence of over 1000 frames. The Video Normals surfaces are registered using the two different regularization algorithms described in Section 5. In this experiment, the first depthmap of the face

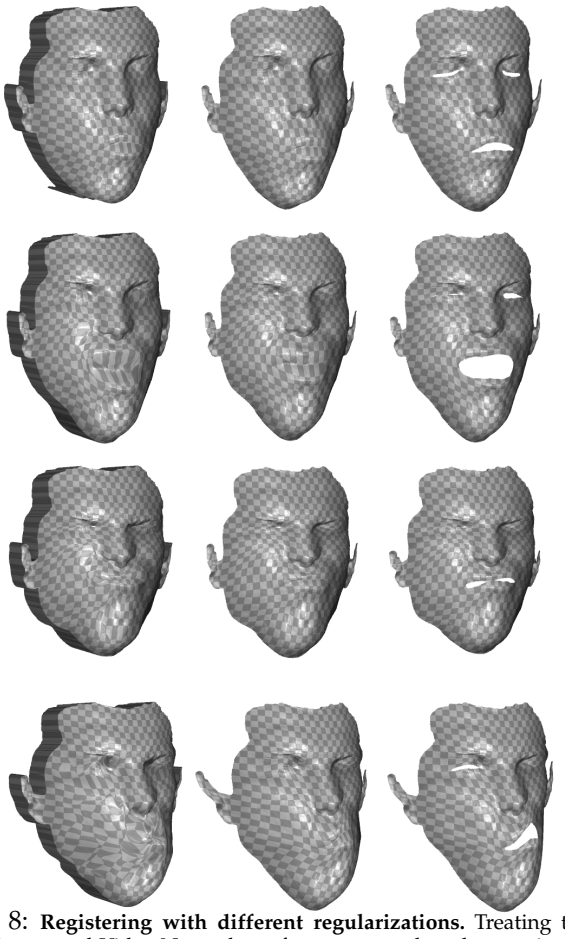


Fig. 8: **Registering with different regularizations.** Treating the first integrated Video Normals surface as a template that receives a checkerboard texture, we automatically register that shape throughout a long sequence by tracking flow frame-to-frame. The rows feature frame #10, #87, #116, and #290 out of 1000. The left column uses the original translation regularization from [14], while the middle column was registered using the alternative Laplacian coordinates regularizer. Results on the right are generated like the middle column, but with the benefit of slits for the mouth and eyes, so some domain-specific input from a user.

sequence is used as a template for the rest of the sequence. The left column of Figure 8 shows the result of using the same rigidity constraint on the translation vectors as for the white sweater, and as described originally in [14]. The performance of this algorithm degenerates most quickly. This is expected, since the face undergoes much bigger deformations than the cloth sequence, so, imposing rigidity on the translation vectors is not enough. The middle column of Figure 8 shows the tracking results using our alternative regularization, the Laplacian coordinates algorithm similar to [63]. This algorithm is better able to impose rigidity constraints. However, the results show the limitations of using optical flow for large deformations. The optical flow easily accumulates errors, and even though rigidity does help in recovering from flow errors, it eventually cannot cope with the amount of deformation shown in this sequence. One possible avenue is to incorporate the work of

Fig. 9: **Cloth reconstruction results of a deforming sweater.** Multispectral photometric reconstruction of a single frame of a longer video sequence using the technique described in Section 3. Multiple viewing angles (frontal, ± 25 degrees, ± 50 degrees) of frame #380 of the sweater sequence. This frame is representative of the detail quality in our reconstructions for this and other tested videos.

[45], though memory limitation hinder this. Their algorithm is also targeted at deforming point clouds, which is a harder problem than ours. Their example results do not exhibit nearly as much deformation as this face sequence. Finally, with human supervision, some of the deformation artifacts due to the eyes and mouth opening and closing can be alleviated by introducing seams on the template at the mouth and eye positions (see Figure 8 right). The seams allow better tracking of large deformations, but the added degrees of freedom can also negatively affect the overall shape. Naturally, eventually, even the rightmost registration accumulates too much error.

7.3 'Dressing' a virtual character with moving cloth

To demonstrate the potential of our method for capturing cloth for animation, we attach a captured moving mesh to an articulated skeleton. Skinning algorithms have varying degrees of realism and complexity, e.g. [68]. We apply a version of smooth skinning in which each vertex $\mathbf{v}_k$ in the mesh is attached to one or more skeleton joints and a link to joint i is weighted by $w_{i,k}$. The weights control how much each joint i affects the transformation of the vertex [69]

$$\mathbf{v}_k^t = \sum_i w_{i,k} \mathbf{S}_i^{t-1} \mathbf{v}_k^{t-1} \quad , \quad \sum_i w_{i,k} = 1, \qquad (8)$$

where the matrix $\mathbf{S}_i^t$ represents the transformation from joint i's local space to world space at time instant t. The mesh is attached to the skeleton by first aligning both in a fixed pose and then finding, for each mesh vertex, a set of nearest neighbors on the skeleton. The weights are set inversely proportional to these distances. The skeleton is animated using publicly available mocap data [70] while the mesh is animated by playing back one of our captured and registered cloth sequences. Figure 11 shows example frames from the rendered sequence (please also see the video). Even though the skeleton and cloth motions are not explicitly aligned, the visual effect of the cloth moving on a controllable character is appealing. Such data-driven cloth animation can serve as a useful tool and presents an alternative to physical cloth simulation.

Fig. 10: **Cloth tracking results of a sweater sequence.** First row: input video sequence of a person wearing a white sweater while being illuminated by three colored lights from three different orientations. Second row: *video of depth-maps* obtained by the technique described in Section 3 and directly texture mapped without any registration. The approach is quickly seen to fail after a few frames. Third row: texture-mapping is obtained by advecting frame-by-frame 2D optical flow [33]. Error in the optical flow advection causes artifacts after about 380 frames. Fourth row: First method of Section 5, where 2D optical flow is regularized with a translational rigidity constraint to reduce advection errors. Last row: on the left and right respectively are close-ups of frame #508 taken from the third and fourth rows. Please see the video.

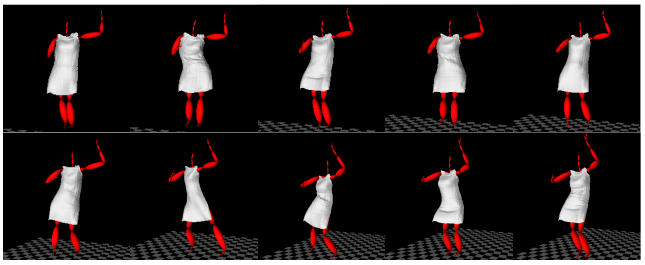


Fig. 11: **Attaching captured moving cloth to an animated character.** We apply smooth skinning to attach a moving mesh to an articulated skeleton that can be animated with mocap data. The mesh is simply animated by playing back the captured and registered dancing cloth sequence (please also see the video).

8 Conclusion

Building on the long established but surprisingly overlooked theory of multispectral lighting for photometric stereo, we have discovered and overcome several new obstacles. We developed a capture methodology that parallels existing work for capturing static cloth, but also enables one to capture the changing shape of cloth in motion. The same technique works well for capturing deforming faces, when the actor wears white makeup. Further, our SfM-based reflectance calibration technique empowers us to compute Video Normals of natural skin color, *without* any makeup. Realtime integration of the resulting normal fields is possible with an FFT normal map integration algorithm using CUDA libraries. We have verified the accuracy of the depth-maps against classic photometric stereo, and measured the space-time accuracy of normals using a rigid but moving shape. When a sequence of reconstructed surfaces is played back, they appear to change smoothly, even under abrupt motions like flutter in strong wind. We also explored long-term registration, and have devised a method to detect and cope with mild self-shadowing.

The high level of detail captured by the normal fields includes surface bends, wrinkles, and even temporary folds. Tracking of folds and parting surfaces like eyelids is inherently underconstrained, and continues to be a challenge, and special templates may help [34], as may other domain-specific constraints about the subject's surface. Instead of [62], [71] could be used to advect flow only in confident areas. Our system could also be extended for some scenes to incorporate the gradual-change prior of [72]. A different mathematical model will need to be explored for non-Lambertian and multi-hue materials. Another limitation is that in-the-round capture would be challenging to arrange, because multiple triples of lights would have to be set up, and they would need to have non-overlapping wavelengths of light. Registration remains the biggest limitation when making use of our monocular capture system, as illustrated in our long sequences. This problem is not singular to Video Normals, so we hope that our shared data proves useful to other researchers as well.

References

[1] J. Pilet, V. Lepetit, and P. Fua, "Real-time non-rigid surface detection," in *Proc. IEEE CVPR*, 2005.

[2] M. Salzmann, S. Ilic, and P. Fua, "Physically valid shape parameterization for monocular 3-d deformable surface tracking," in *British Machine Vision Conference*, 2005.

[3] V. Scholz, T. Stich, M. Keckeisen, M. Wacker, and M. Magnor, "Garment motion capture using color-coded patterns," *Computer Graphics Forum (Proc. Eurographics EG'05)*, vol. 24, no. 3, pp. 439–448, Aug. 2005.

[4] R. White and D. Forsyth, "Retexturing single views using texture and shading," in *ECCV*, 2006, pp. 70–81.

[5] R. White, K. Crane, and D. Forsyth, "Capturing and animating occluded cloth," in *ACM Trans. on Graphics (SIGGRAPH)*, 2007.

[6] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 519–528.

[7] A. Hertzmann and S. Seitz, "Shape and materials by example: a photometric stereo approach," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003, pp. I: 533–540.

[8] C. Hernández, G. Vogiatzis, and R. Cipolla, "Multiview photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 548–554, 2008.

[9] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk, "The Digital Michelangelo Project: 3d scanning of large statues," in *Proc. of the ACM SIGGRAPH*, 2000, p. 1522.

[10] T. Malzbender, D. G. B. Wilburn, and B. Ambrisco, "Surface enhancement using real-time photometric stereo and reflectance transformation," in *Eurographics Symp. on Rendering 2006*, 2006.

[11] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, 2008.

[12] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," *ACM Trans. Graph.*, vol. 27, no. 3, 2008.

[13] C. Hernández and G. Vogiatzis, "Self-calibrating a real-time monocular 3d facial capture system," *Intl. Symp. 3DPVT*, 2010.

[14] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla, "Non-rigid photometric stereo with colored lights," in *Proc. of the 11th IEEE Intl. Conf. on Comp. Vision (ICCV)*, 2007.

[15] D. Pritchard and W. Heidrich, "Cloth motion capture," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 263–272, 2003.

[16] R. Woodham, "Photometric method for determining surface orientation from multiple images," in *Optical Eng.*, vol. 19, no. 1, 1980, pp. 139–144.

[17] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman, "Passive photometric stereo from motion," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, 2005, pp. 1635–1642.

[18] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and spatially-varying BRDFs from photometric stereo," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol. 1, 2005, pp. 341–348.

[19] A. Hertzmann and S. Seitz, "Shape reconstruction with general, varying BRDFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1254–1264, November 2005.

[20] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *EGSR*, 2007.

[21] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," *ACM ToG (Proc. SIGGRAPH Asia)*, vol. 28, no. 5, 2009.

[22] 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," in *ACM Trans. Graph.*, 2008, pp. 1–10.

[23] C. A. Wilson, A. Ghosh, P. Peers, J.-Y. Chiang, J. Busch, and P. Debevec, "Temporal upsampling of performance geometry using photometric alignment," *ACM Trans. Graph.*, vol. 29, no. 2, pp. 1–11, 2010.

[24] A. Petrov, "Light, color and shape," *Cognitive Processes and their Simulation (in Russian)*, pp. 350–358, 1987.

[25] L. Kontsevich, A. Petrov, and I. Vergelskaya, "Reconstruction of shape from shading in color images." *J. Opt. Soc. Am. A*, vol. 11, no. 3, pp. 1047–1052, 1994.

[26] M. S. Drew and L. L. Kontsevich, "Closed-form attitude determination under spectrally varying illumination," in *Proc. CVPR*, 1994, pp. 985–990.

[27] R. J. Woodham, "Gradient and curvature from the photometric-stereo method, including local confidence estimation," *J. Opt. Soc. Am. A*, vol. 11, no. 11, pp. 3050–3068, 1994.

[28] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1070–1077.

[29] K. S. Bhat, C. D. Twigg, J. K. Hodgins, P. K. Khosla, Z. Popović, and S. M. Seitz, "Estimating cloth simulation parameters from video," in *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer animation*, 2003, pp. 37–51.

[30] X. Gu, S. Zhang, P. Huang, L. Zhang, S.-T. Yau, and R. Martin, "Holoimages," in *SPM '06: Proc. 2006 ACM Symposium on Solid and Physical Modeling*. ACM Press, 2006, pp. 129–138.

[31] T. Weise, B. Leibe, and L. V. Gool, "Fast 3d scanning with automatic motion compensation," in *Proc. IEEE CVPR '07*, 2007.

[32] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: High-resolution capture for modeling and animation," in *ACM Annual Conf. on Computer Graphics*, 2004, pp. 548–558.

[33] P. Sand, L. McMillan, and J. Popović, "Continuous capture of skin deformation," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 578–586, 2003.

[34] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM ToG (Proceedings SIGGRAPH Asia 2009)*, vol. 28, no. 5, 2009.

[35] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur, "Markerless garment capture," *ACM ToG*, vol. 27, no. 3, 2008.

[36] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE CG&A*, vol. 27, no. 3, pp. 21–31, 2007.

[37] N. Ahmed, C. Theobalt, P. Dobre, H.-P. Seidel, and S. Thrun, "Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry," in *IEEE CVPR*, 2008, pp. 1–8.

[38] Y. Pekelny and C. Gotsman, "Articulated object reconstruction and markerless motion capture from depth video," *Computer Graphics Forum*, vol. 27, no. 2, pp. 399–408, Apr. 2008.

[39] W. Chang and M. Zwicker, "Automatic registration for articulated shapes," *Computer Graphics Forum (Proceedings of SGP 2008)*, vol. 27, no. 5, pp. 1459–1468, 2008.

[40] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 433 – 449, May 1999.

[41] A. Tagliasacchi, H. Zhang, and D. Cohen-Or, "Curve skeleton extraction from incomplete point cloud," *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2009)*, vol. 28, no. 3, pp. Article 71, 9 pages, 2009.

[42] Q. Zheng, A. Sharf, A. Tagliasacchi, B. Chen, H. Zhang, A. Sheffer, and D. Cohen-Or, "Consensus skeleton for non-rigid space-time registration," *Computer Graphcis Forum (Special Issue of Eurographics)*, vol. 29, no. 2, pp. 635–644, 2010.

[43] A. Sharf, D. A. Alcantara, T. Lewiner, C. Greif, A. Sheffer, N. Amenta, and D. Cohen-Or, "Space-time surface reconstruction using incompressible flow," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–10, 2008.

[44] N. J. Mitra, S. Flory, M. Ovsjanikov, N. Gelfand, L. Guibas, and H. Pottmann, "Dynamic geometry registration," in *Computer Graphics Forum (Proceedings of SGP 2007)*, 2007, pp. 173–182.

[45] J. Süßmuth, M. Winter, and G. Greiner, "Reconstructing animated meshes from time-varying point clouds," *Computer Graphics Forum (Proceedings of SGP 2008)*, vol. 27, no. 5, pp. 1469–1476, 2008.

[46] M. Wand, B. Adams, M. Ovsjanikov, A. Berner, M. Bokeloh, P. Jenke, L. Guibas, H.-P. Seidel, and A. Schilling, "Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data," *ACM. Trans. Graph.*, vol. 28, no. 2, p. 15, Apr. 2009.

[47] M. S. Drew, "Direct solution of orientation-from-color problem using a modification of Pentland's light source direction estimator," *Comput. Vis. Image Underst.*, vol. 64, no. 2, pp. 286–299, 1996.

[48] J. A. Paterson, D. Claus, and A. W. Fitzgibbon, "BRDF and geometry capture from extended inhomogeneous samples using flash photography," *Computer Graphics Forum (Special Eurographics Issue)*, vol. 24, no. 3, pp. 383–391, 2005.

[49] T. Simchony, R. Chellappa, and M. Shao, "Direct analytical methods for solving poisson equations in computer vision problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 435–446, 1990.

[50] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," *ACM Trans. on Graphics (Proc. SIGGRAPH)*, vol. 29, no. 3, 2010.

[51] Y. Furukawa and J. Ponce, "Dense 3d motion capture for human faces," *IEEE CVPR*, pp. 1–8, 2009.

[52] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," *ACM Trans. on Graphics (Proc. SIGGRAPH)*, vol. 29, no. 3, 2010.

[53] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2004.

[54] Boujou, *2d3 Ltd.*, http://www.2d3.com, 2009.

[55] S. N. Sinha, M. Pollefeys, and L. McMillan, "Camera network calibration from dynamic silhouettes," *Proc. IEEE CVPR*, pp. 195–202, 2004.

[56] C. Hernández, F. Schmitt, and R. Cipolla, "Silhouette coherence for camera calibration under circular motion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 343–349, 2007.

[57] C. Hernández and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding, special issue on 'Model-based and image-based 3D Scene Representation for Interactive Visualization'*, vol. 96, no. 3, pp. 367–392, 2004.

[58] G. Vogiatzis, C. Hernández, P. H. S. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2241–2246, 2007.

[59] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE CVPR*, pp. 1–8, 2007.

[60] C. Hernández, G. Vogiatzis, and R. Cipolla, "Shadows in three-source photometric stereo," in *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, 2008, pp. 290–303.

[61] B. Allen, B. Curless, and Z. Popović, "Articulated body defromation from range scan data." in *Proc. of the ACM SIGGRAPH*, 2002, pp. 612–619.

[62] M. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise smooth flow fields," in *Computer Vision and Image Understanding*, vol. 63(1), 1996, pp. 75–104.

[63] N. Ahmed, C. Theobalt, C. Rossl, S. Thrun, and H. Seidel, "Dense correspondence finding for parametrization-free animation reconstruction from video," 2008, pp. 1–8.

[64] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 213–230, 2008.

[65] P. Sinha and E. H. Adelson, "Recovering Reflectance and Illumination in a World of Painted Polyhedra," in *Proceedings IEEE International Conference on Computer Vision*, 1993, pp. 156–163.

[66] R. Onn and A. Bruckstein, "Integrability disambiguates surface recovery in two-image photometric stereo," *Int. J. Comput. Vision*, vol. 5, no. 1, p. 105, 1990.

[67] R. T. Frankot and R. Chellappa, "A method for enforcing integrability in shape from shading algorithms." *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 439–451, 1988.

[68] J. P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation," in *Proc. ACM SIGGRAPH 2000*, 2000, pp. 165–172.

[69] X. C. Wang and C. Phillips, "Multi-weight enveloping: least-squares approximation techniques for skin animation," in *SCA '02: Proc. of the 2002 ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, 2002, pp. 129–138.

[70] "CMU Graphics Lab Motion Capture Database," http://mocap.cs.cmu.edu.

[71] O. Mac Aodha, G. J. Brostow, and M. Pollefeys, "Segmenting video into classes of algorithm-suitability," in *CVPR*, 2010.

[72] T. Popa, I. South-Dickinson, D. Bradley, A. Sheffer, and W. Heidrich, "Globally consistent space-time reconstruction," *Computer Graphics Forum (Proc. SGP)*, 2010.

Gabriel Brostow received his Ph.D. and M.S. in Computer Science from the Georgia Institute of Technology in 2004, and his B.S. in Electrical Engineering from the University of Texas at Austin in 1996. He was a Marshall Sherfield Fellow and postdoctoral researcher at the University of Cambridge until 2007, and a research scientist at ETH Zurich until 2009. In 2008 he joined the Computer Science Department at University College London as an Assistant Professor. His research focus is "smart capture" of visual data.

Carlos Hernández obtained an MS in applied Mathematics from l'Ecole Normale Supérieure de Cachan in 2000 and received his PhD in 2004 from the Ecole Nationale Supérieure des Télécommunications de Paris. He then moved to the University of Cambridge as a research associate until 2006, when he became a permanent researcher at Toshiba Research Europe. He started working at Google in 2010. His research interests are mainly in shape-from-X algorithms and their applications for Computer Graphics.

George Vogiatzis obtained an MS in Mathematics and Computer Science from Imperial College in 2002 and a PhD in Computer Vision from the University of Cambridge in 2006. He then spent three years in Toshiba Research after which he joined Aston University as a Lecturer. His research interests are in 3D Computer Vision and its applications for Computer Graphics and Animation.

Björn Stenger received the Diplom-Informatiker degree from the University of Bonn in 2000 and the PhD degree from the University of Cambridge in 2004. From 2004-06 he was a Toshiba Fellow at the Toshiba Corporate R&D Center in Kawasaki, Japan, where he worked on gesture recognition and 3D body tracking. He joined Toshiba Research Europe in 2006, where he is currently leading the Computer Vision Group.

Roberto Cipolla received the B.A. degree in engineering from the University of Cambridge, Cambridge, U.K., in 1984, the M.S.E. degree in electrical engineering from the University of Pennsylvania in 1985, and the D.Phil. degree (computer vision) from the University of Oxford, Oxford, U.K., in 1991.

His research interests are in computer vision and robotics and include the recovery of motion and 3D shape of visible surfaces from image sequences; object detection and recognition; novel man-machine interfaces using hand, face and body gestures; real-time visual tracking for localisation and robot guidance; applications of computer vision in mobile phones, visual inspection and image-retrieval and video search. He has authored 3 books, edited 8 volumes and co-authored more than 300 papers.