Robust Fusion of Dynamic Shape and Normal Capture for High-quality Reconstruction of Time-varying Geometry

Naveed Ahmed	Christian Theobalt	Petar Dobrev	Hans-Peter Seidel	Sebastian Thrun
MPI Informatik	Stanford University	Jacobs University Bremen	MPI Informatik	Stanford University

Abstract

This paper describes a new passive approach to capture time-varying scene geometry in large acquisition volumes from multi-view video. It can be applied to reconstruct complete moving models of human actors that feature even slightest dynamic geometry detail, such as wrinkles and folds in clothing, and that can be viewed from 360°. *Starting from multi-view video streams recorded under calibrated lighting, we first perform marker-less human motion capture based on a smooth template with no high-frequency surface detail. Subsequently, surface reflectance and time-varying normal fields are estimated based on the coarse template shape. The main contribution of this paper is a new statistical approach to solve the non-trivial problem of transforming the captured normal field that is defined over the smooth non-planar 3D template into true 3D displacements. Our spatio-temporal reconstruction method outputs displaced geometry that is accurate at each time step of video and temporally smooth, even if the input data are affected by noise.*

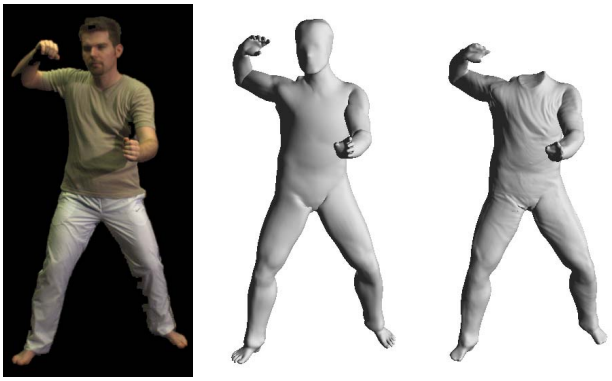
1. Introduction

For creating high quality animations of real world scenes in a computer, it is essential that accurate models of shape and appearance are at hand. Hand-crafting detailed moving scene geometry is a cumbersome process, as it requires tedious manual work or computationally expensive numerical simulations (e.g. for clothing). The development of scanning devices that deliver fine-grained shape models of at least static scenes has therefore greatly facilitated animation production. Unfortunately, capturing high-quality time-varying shape of dynamic scenes at the same level of fidelity is still a big challenge. First approaches to reach this goal were based on active video-based measurement, such as structured light, or employed a combination of visual hull and stereo. While the former approaches are merely usable for small-scale scenes (e.g. faces) and interference makes multi-view recording difficult, stereo approaches often fall short in delivering the high level of accuracy that computer animation requires (Sect. 2).

In contrast, we propose a new method to passively capture highly-detailed dynamic surface geometry of humans from multiple video recordings under calibrated lighting. Our algorithm capitalizes on and extends the ideas originally proposed in the work by Theobalt et al. [19]. In their original work they first perform marker-less motion capture on the input data in order to make a coarse kinematic template (shown in Fig. 1b) follow the motion of the actor. Subsequently, they reconstruct a reflectance model for each point on the surface, and exploit this knowledge to measure a dynamic surface normal field parametrized over the smooth template. While this representation was sufficient for their relightable 3D video rendering application, they did not approach the difficult problem of converting a potentially noise-contaminated normal field parametrized over an arbitrarily shaped smooth surface into highly-detailed time-varying scene geometry.

Our paper allows us to do exactly the latter. Our first contribution is an improvement over Theobalt et al.'s original surface reflectance and normal estimation approach which now employs robust statistics to handle sensor noise more faithfully, Sect. 3.2. Our second and most important contribution is a new spatio-temporal deformation framework that enables us to transform the moving template geometry and the time-varying normal field into true spatio-temporally varying scene geometry that reproduces geometric surface detail at millimeter-scale accuracy. Standard normal field integration schemes are not feasible in this setting as they often perform poorly in the presence of noise and as they do not easily generalize to the case of arbitrarily oriented base surfaces in 3D. In contrast, we formulate the problem as a spatio-temporal Markov Random field such that we can reconstruct fine-grained geometry that is spatially accurate, as well as temporally smooth, even if the input was affected by noise.

We demonstrate and validate the accuracy of our method based on several sequences that were kindly provided to us by the authors of [19], Sect. 5.



(a) (b) (c)

Figure 1. Input video frame (a), smooth 3D template model in same pose (b), our detailed 3D surface model with true geometric detail such as wrinkles on the shirt (c).

2. Related Work

Nowadays, most systems that can capture dynamic scene geometry at millimeter scale accuracy are restricted to confined spatial volumes, e.g. structured light systems for facial performance capture [22]. Mainly due to interference and spatial resolution issues, it is hard to apply these methods for capturing humans from multiple views. While a combination of shape-from-silhouette and stereo is one way to approach the latter scenario, the inherent difficulty and lack of robustness in stereo make it hard to achieve very high accuracy and resolution [6, 18].

An alternative to normal stereo in a static case which has the potential to capture fine-grained surface detail is photometric stereo, a variant of shape-from-shading. In photometric stereo one makes assumptions about surface reflectance properties to recover normal orientation from images taken under varying lighting [20, 23]. It has also been tried to simultaneously estimate reflectance (e.g. BRDF information) and normal data from a variety of 2D images which were taken under calibrated lighting [9, 10]. In this single 2D view case, normal field integration schemes can be applied to transform orientation data into true highly-detailed height values [7, 3]. Chang et. al [5] used level set methods to integrate multi-view normal fields.

While it is feasible to estimate BRDF and normal orientation also for more general static 3D objects that were photographed under a variety of viewpoints and light directions [16], the deformation of geometry based on normals parametrized over a general 3D shape is non-trivial. Standard integration schemes (assuming orthographic projection and height fields that are parametrized over a plane) are not applicable anymore since absolute 3D position has to be recovered and coherence of the displacements over non-planar geometry needs to be assured.

One way to attack this problem is to measure 3D position approximately, e.g. by stereo or structured light scanning, and use normal information obtained via shape from shading to improve the initial position estimates and the degree of surface detail [11]. While early work in this direction produced comparably coarse 3D geometry [8, 15], the work by Nehab et al. [17] produces detailed models of static objects by refining scanned 3D point positions until photometrically measured normals are well approximated. Jones et al. [13] applied the latter technique to improve captured dynamic face geometry, but they did not formulate it as a spatio-temporal problem nor does their setup scale easily to larger scenes.

We capitalize on this idea as well but develop a more advanced reconstruction approach suitable for large-scale dynamic scenes. In contrast to previous work, our approach generates geometry that is accurate and detailed at each time step, *and* that is coherently deforming over time. We also incorporate characteristics of measurement noise into the reconstruction process by posing our problem as a spatio-temporal Markov Random Field (MRF).

Our starting point is the work and data by Theobalt et al. [19] who capture shape, motion, reflectance and time-varying normals of human actors from only a handful of synchronized video recordings under calibrated lighting. Their method parametrizes shape motion and reflectance based on a smooth template body model that lacks any geometric detail. In this paper, we improve their reflectance and normal field estimation approach by using robust statistics. We then propose a new spatio-temporal MRF framework which transforms smooth geometry and normals into highly detailed dynamic scene geometry even in the presence of notable measurement noise. As we can process normal fields over arbitrarily shaped time-varying base surfaces in 3D we can capture time-varying geometry at detail levels unparalleled by any related approach, such as purely stereo-based reconstruction methods mentioned earlier.

3. Problem Statement

Our goal is to passively reconstruct accurate and highly detailed dynamic surface geometry of humans from only a handful of synchronized video recordings, Sec. 3.1 and Fig. 2. To this end, the motion of the actor in input video recordings is tracked by means of the marker-less motion estimation approach presented in [19]. This method parametrizes dynamic scene geometry in the form of an adaptable kinematic body template with smooth surface geometry (the tracked geometry for all our test data was kindly provided to us by the authors of the original work). If the original video sequences were recorded under calibrated lighting, surface reflectance properties, i.e. per-surface-point BRDfs, as well as dynamic normal maps can be estimated as shown in [19]. We pick up and extend the ideas

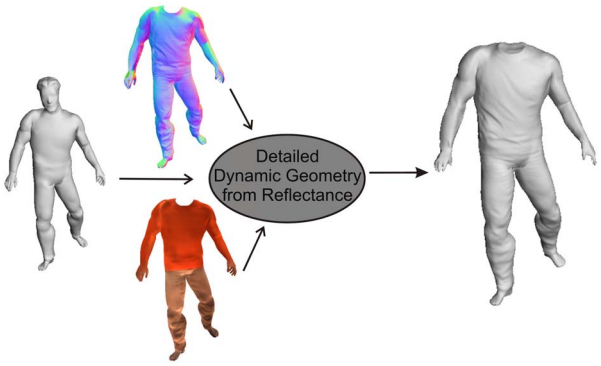


Figure 2. Overview: The tracked smooth template model (left), along with per-texel refined normal field (top) and per-texel BRDF parameters (bottom) are used to estimate detailed time-varying surface geometry (right).

presented in their original paper such that we can use the same acquisition setup to acquire dynamic geometry at an unprecedented detail level. We demonstrate our method on sequences that were kindly provided to us by the authors of the original work.

To achieve our goal, we first modify the original BRDF estimation pipeline by including robust statistics into the reconstruction framework, Sect. 3.2. Thereafter, we estimate dynamic normal (bump maps) from the input video sequences that are defined over the smooth template geometry, Sect. 3.3. Finally, we develop a spatio-temporal Markov-Random-Field-based surface refinement procedure which is one of the first to enable integration of normal fields on arbitrarily shaped time-varying template geometry. Our new spatio-temporal framework captures at the same time spatially accurate and temporally smooth geometry and handles sensor noise robustly, Sect. 4.

3.1. Data Acquisition and Template Motion Estimation

The basics we can capitalize on, i.e. the acquisition procedure, the employed template model and the marker-less motion estimation approach have been described in detail in [19]. In this and the following subsections, we briefly recapitulate the original approach for ease of understanding, and point out some important modifications we made.

Input sequences are recorded from eight static video cameras running at 1004x1004 pixels resolution, 25 fps, and 12 bit color resolution. the cameras are arranged in an approximately circular setup around the center of the scene. The whole scene background is draped in black molleton and the only light sources in the scene are two calibrated HMI lamps placed in opposite corners of the acquisition space. For each person and each set of clothing, two types of sequences are captured. In the so-called reflectance estimation sequence (RES), the person strikes a static body pose and slowly rotates on spot. From this sequence, a BRDF model for each surface point is estimated. The second type of sequence, the dynamic scene sequence (DSS), shows arbitrary motion of the person and is used for dynamic normal map and dynamic geometry reconstruction.

The shape and motion of the actor is parametrized by means of a template model comprising of a kinematic skeleton, as well as a closed triangle mesh $\mathcal{M}$ that represent surface geometry, Fig. 2. In a preprocessing step the model is scaled and deformed to match the outline of the recorded actor. A marker-less silhouette-based motion estimation approach is employed to capture the pose of the actor at each RES and DSS frame, yielding a sequence of configurations of $\mathcal{M}$ in T poses $\mathcal{M}(t), t = 1, \ldots, T$.

Prior to reconstruction, $\mathcal{M}$ is parametrized over a 2D square by means of the conformal mapping technique described in [21]. To this end, the closed mesh is manually cut open during pre-processing to obtain a free boundary. In the following, we refer to parametrized positions on the template surface as $u_{i,j}$, and to the corresponding positions in 3D space at time t as $\mathbf{x}(u_{i,j}, t)$. Based on this parametrization, all input camera images are transformed into the texture domain. Prior to texture image generation, an image-based texture warping approach is applied in order to correct misregistrations due to shifting apparel, as well as due to mismatches between template and true geometry [19].

3.2. BRDF Estimation

After performing marker-less motion capture for each frame of multi-view video, the position and orientation of each $u_{i,j}$ with respect to the calibrated acquisition apparatus is known. In other words, due to the scene motion it becomes possible to collect for each point on the surface a variety of reflectance samples, each representing the appearance of the point from known outgoing viewing and incoming lighting directions. The method described in [19] exploits this fact in order to estimate for each $u_{i,j}$ a static parametric BRDF model from the RES.

In their original approach, an energy minimization framework was used to compute parameters of an isotropic Lafortune BRDF f_r at each surface point such that the measured data are best approximated [14]. In our research, we replace their original least-squares approach by a regression framework based on robust Huber statistics [12] as this enables us to obtain more faithful estimates in the presence of non-Gaussian measurement noise. For each surface point $u_{i,j}$ on the template, we minimize the following energy functional to find an isotropic BRDF that reproduces the

data in the RES:

$$E_{\text{BRDF}}(\rho(u_{i,j})) = \sum_{t}^{T}\sum_{c}^{8}\kappa_c(u_{i,j},t)\mathcal{H}\Big(S_c(u_{i,j},t) - [\sum_{e}^{2}\lambda_e(u_{i,j},t)(t)(f_r(\mathbf{l}(u_{i,j},t),\mathbf{v}_c(u_{i,j},t),\rho(u_{i,j})) \cdot I_e(\mathbf{n}_o(u_{i,j},t)\cdot\mathbf{l}(u_{i,j},t)))]\Big)^2. \quad (1)$$

E_{BRDF} is evaluated separately in the red, green and blue color channel. $S_c(u_{i,j},t)$ denotes the color of $u_{i,j}$ measured from camera c, and I_e denotes the intensity of light source e. The viewing directions $\mathbf{v}_c(u_{i,j},t)$ and light source directions $\mathbf{l}_e(u_{i,j},t)$ are expressed in $u_{i,j}$'s local coordinate frame based on the (template) surface normal $\mathbf{n}_o(u_{i,j},t)$. $\kappa_c(u_{i,j},t)$ and $\lambda_e(u_{i,j},t)$ encode the visibility of point $u_{i,j}$ with respect to cameras and light sources, respectively. As opposed to the original least-squares minimization framework of [19] which assumes Gaussian noise in reflectance samples and thus may over-weight outliers, we employ robust Huber statistics $\mathcal{H}$ as penalizer, see Appendix. $\mathcal{H}$ preserves the advantageous convergence properties of an L_2 function for inliers, but resorts to an L_1 norm for samples that are likely to be outliers. By this means we implicitly model our noise characteristics more faithfully as a heavy-tail Gaussian. In order to find the clip threshold k for $\mathcal{H}$ we analyze the variance in captured reflectance samples in a series of consecutive video frames in which the person remains in a static pose relative to the cameras. For each color channel and each material we compute the average variance and use the squared values as material- and color-specific clip thresholds.

In practice BRDF parameters are estimated in a multi-step procedure. First, materials on the surface are clustered based on average diffuse color and a specular BRDF component is estimated for each material separately. Thereafter, a per-texel diffuse model is fit to each surface point after subtracting the previously estimated specular component from each sample. Please note that we only use samples seen by exactly one light source for estimation, which, due to the positioning of lamps in the studio, in reality is true for over 90% of samples. For numerical minimization, we employ the L-BFGS-B minimizer [4].

3.3. Dynamic Normals from Reflectance

Given the detailed BRDF data estimated from the RES, a variant of shape from shading can be employed to compute an accurate normal map for the whole template at each time step of a DSS sequence. Despite being parametrized over the smooth template this normal map contains information on true time-varying geometric detail in the form of a direction field. To compute a new normal direction $\mathbf{n}_m(u_{i,j},t)$ at each time step, the following energy is minimized for each surface location:

$$E_{\text{normal}}(\mathbf{n}_m(u_{i,j},t)) = \omega E_{\text{BRDF}}(\rho(u_{i,j})) + \mu\Delta(\mathbf{n}_m(u_{i,j},\mathbf{n}_o(u_{i,j})))^{\epsilon}. \quad (2)$$

Here, E_{BRDF} is the original BRDF error term instantiated with the previously computed BRDF parameters. Unfortunately, the problem of solving for the normal direction by only considering samples from a single video frame - i.e. a single light source position - is ill-conditioned. Therefore, we make the assumptions that the normal direction in a local frame does not change in a small time interval, and solve for a constant normal direction over a small time window of subsequent frames (typically 5) in which $u_{i,j}$ has been seen under different light directions. Finally, normal directions are interpolated in the time domain.

To further regularize our solution, we add an additional term $\Delta(\mathbf{n}_m(u_{i,j},t),\mathbf{n}_o(u_{i,j},t))^{\epsilon}$ that penalizes deviations Δ of the measured normals from the original normals of the smooth template. The penalty exponent ϵ, as well as the importance weights ω and μ that sum to 1, are found through experiments.

4. Adding Spatio-temporally Coherent Geometric Surface Detail

Dynamic normal fields encode information on high-frequency surface detail without physically deforming the smooth template surface over which they are parametrized. This information is sufficient to render relightable 3D videos of humans from many angles apart from [19] grazing ones. However, true 3D time-varying geometric detail is essential in many production quality animation settings where full global illumination renderings are expected. Only true deformed surface geometry will enable correct appearance of the shape under the final lighting simulation.

In the following, we therefore present a new data fusion framework that transforms the original setup for relightable 3D video capture into a system for high-quality capture of detailed dynamic surface geometry. Our method is grounded on the assumption that our smooth template, essentially capturing low frequency geometry, is already well-aligned with the input.

Our algorithm estimates for each surface point $\mathbf{u}_{i,j}$ on the smooth template at each time step t a 3D displacement vector $\mathbf{d}(u_{i,j},t)$ that yields the true 3D position of the point u at t as $\mathbf{x}_d(u_{i,j},t) = \mathbf{x}(u_{i,j},t) + \mathbf{d}(u_{i,j},t)$. Since the true displacements are expected to be small, it is safe to assume that the displacement direction is always along the direction of template normals.

As our measurements are potentially contaminated by noise, we employ a statistical framework to robustly find the most likely field of surface displacements given the data. To achieve this purpose we model the joint posterior distribution of the field of displacements at each time step as a Markov Random Field (MRF) which takes the form

$$\mathbf{p}(\mathbf{d}(u_{i,j}t,) \,|\mathbf{n}_m(u_{i,j},t),\mathcal{M}(t)) = \frac{1}{Z}e^{-(\alpha\Phi(t)+\beta\Psi(t)+\gamma\Omega(t)+\delta\Xi(t))} \,, \quad (3)$$

where Z is a normalization constant, $\Phi(t)$ models our measurement process, and $\Psi(t)$, $\Omega(t)$ and $\Xi(t)$ are prior potentials. α,β, γ and δ are weighting factors summing to 1. Empirically we found that values of $\alpha = 0.6$, $\beta = 0.1$, $\gamma = 0.2$ and $\delta = 0.1$ produce most decent results (see also Sect. 5 for a discussion). The spatio-temporal neighborhood structure of our MRF connects each surface location to the four immediate spatially adjacent ones at the same time step (easily found from our surface parametrization), as well as to its instantiations at the two previous time steps.

As we are interested in the most likely solution given the current data only and not in the full posterior, we find the most likely surface as the maximum a posteriori (MAP) hypothesis by minimizing the negative log-likelihood of (3) as

$$\hat{\mathbf{d}}(u_{i,j},t) = \underset{\mathbf{d}(u_{i,j},t)}{\operatorname{argmin}} \; \alpha\Phi(t) + \beta\Psi(t) + \gamma\Omega(t) + \delta\Xi(t) \,. \quad (4)$$

In the following subsections, we first describe and motivate how assumptions about noise characteristics are encoded in measurement potentials, Sect. 4.1, and illustrate what prior potentials are appropriate to properly condition our solution space, Sect. 4.2. Finally, we describe how to practically solve for a maximum a posteriori (MAP) surface even in our large scenes with on average 350,000 surface points, Sect. 4.3.

4.1. Measurement Potential

The information that captures the true shape of the fine-grained surface details is encoded in our measured surface normal field $\mathbf{n}_m(u,t)$. Our measurement potential therefore aims at minimizing the angular difference $\Delta(\mathbf{n}_m(u_i,t),\mathbf{n}_r(u_i,t))$ between the measured normals and the normals of the displaced surface.

To properly constrain our problem, we don't formulate the error in normal field approximation based on individual locations $u_{i,j}$ (i.e. individual texels in the texture domain), but rather based on triangles obtained by regularly triangulating all texels in the parametrization. Normals for the obtained triangles are computed by simply averaging the normals at its three vertices (i.e. texels). Again, we capitalize on the Huber function $\mathcal{H}$ to obtain more reliable estimates in the presence of noise. Our measurement potential thus takes the form

$$\Phi(t) = \sum_{D=(u_a,u_b,u_c)\in\mathbf{D}} \mathcal{H}(\Delta(\mathbf{n}_m(D,t),\mathbf{n}_r(D,t))) \,, \quad (5)$$

where $D = (u_a, u_b, u_c)$ is a triangle formed by adjacent texels (surface points) u_a, u_b, and u_c, and $\mathbf{D}$ is the set of all such triangles. $\mathbf{n}_r(D,t)$ is the normal field according to the current deformed surface evaluated at D, and $\mathbf{n}_m(D,t)$ is the respective measured normal field. The clip threshold k was chosen conservatively in such that deviations of new and measured normals by more than $90°$ are considered outliers.

4.2. Prior Potentials

We make the general assumption that dynamic surfaces in the real world are smooth in both space and time. In other words, spatially adjacent surface locations should exhibit similar displacements and the change in displacement for the same surface location over time should be in reasonable bounds as well. The spatial smoothness constraint penalizes local deviation from an oriented plane in a 4-neighborhood around each point and is encoded in the potential

$$\begin{aligned}\Psi(t) = \sum_i \sum_j & \mathcal{H}(\mathbf{x}_d(u_{i-1,j},t) - 2\mathbf{x}_d(u_{i,j},t) + \\ & \mathbf{x}_d(u_{i+1,j},t)) + \\ & \mathcal{H}(\mathbf{x}_d(u_{i-1,j},t) - 2\mathbf{x}_d(u_{i,j},t) + \\ & \mathbf{x}_d(u_{i+1,j},t)) \,,\end{aligned} \quad (6)$$

where $\mathbf{x}_d(u_{i-1,j},t)$, $\mathbf{x}_d(u_{i+1,j},t)$, $\mathbf{x}_d(u_{i,j-1},t)$, and $\mathbf{x}_d(u_{i,j+1},t)$ are displaced 3D positions of surface locations adjacent to $u_{i,j}$. The clip threshold k of $\mathcal{H}$ in this case is chosen such that differences in local surface normal orientation of more than $30°$ are considered outliers.

Temporal smoothness is enforced by the potential

$$\begin{aligned}\Xi(t) = \sum_i \sum_j & (\mathbf{d}(u_{i,j},t) - 2\mathbf{d}(u_{i,j},t-1) - \\ & \mathbf{d}(u_{i,j},t-2))^2 \,.\end{aligned} \quad (7)$$

This term favors a smooth rate of change of displacements over time, or putting it differently, favors small "acceleration" in displacement change over time.

Lastly, we make the a priori assumption that displaced surface locations should remain close to the original smooth template shape. The latter constraint is essential as it prevents our surface to drift arbitrarily far away from the orig-

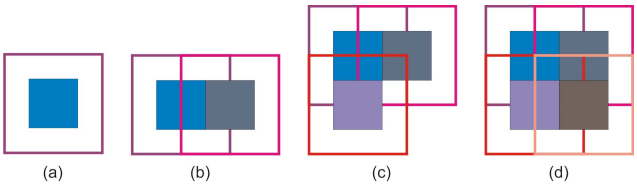


Figure 3. Patch-based optimization. A single patch, its boundary area, and its (blue) internal area (a). While the deformed surface is computed, the overlapping patches are processed in a sequence as shown in (b), (c) and (d) respectively. Only the interior patch area is preserved after displacement computation for one patch.

inal template. Our second prior therefore takes the form

$$\Omega(t) = \sum_i \sum_j \mathbf{d}(u_{i,j}, t)^2 \tag{8}$$

4.3. Practical Implementation

The test sequences provided to us by Theobalt et al. feature parametrizations of the smooth template of size 1024×1024 pixels. On average this corresponds to $350,000$ surface locations for which a displacement needs to be found at each time step. Please note that we compute displacements at a much higher level of granularity than the vertex density of the original template which is typically only 40,000. Parametrizations were obtained by manually cutting the template open and unfolding it over a 2D square by means of the conformal mapping technique described in [21].

As we are only interested in a MAP solution to the final surface, we can conveniently resort to standard off-the-shelf L-BFGS-B technique [4] to minimize(4).

To keep optimization tractable in the light of our very dense surface sampling, we also subdivide the overall surface reconstruction problem into a series of smaller ones. In practice, we subsequently compute displacements for individual surface patches and successively merge information from different patches to create the final result.

Each patch on our model corresponds to a square region of surface locations in our parametrization domain. Furthermore, each such square region is composed of an interior region and an exterior boundary area, Fig. 3. If we would simply deform individual adjacent patches we would with very high likelihood obtain discontinuities at patch boundaries since the mutual MRF dependencies across the rim are not properly considered. To prevent this source of error, we arrange subsequent patches in an interleaving, half-overlapping pattern, see Fig. 3b,c,d for the temporal sequence in which the patches are processed. Furthermore, after the displacements for one complete patch were estimated, we only preserve the displacement at the center of the patch. The boundary regions are thus only employed to initialize the optimization of any subsequent patch whose center region overlaps with the boundary. All patches are considered equal, thus the choice of the starting patch for our optimization is arbitrary. Overall, this interleaved optimization pattern produces a high quality surface estimate that preserves detail while preventing erroneous discontinuities along boundaries, see Sect. 5 for further discussion.

5. Results and Validation

To demonstrate the results of our method, we have used two captured real-world motion sequences that the authors of [19] provided us. The data for each sequence comprises of the moving low-detail template, all input image data (also in texture format already), full calibration data (cameras and lights), parametrization and warp-corrected texture coordinates. The latter is a set of data which encodes information on cloth shifting over the body's surface which was detected by a method detailed in [19].

The first sequence shows a scene in which the actor wears mostly diffuse clothing and walks back and forth in front of the cameras, Fig 5a,b. The RES (used for BRDF estimation) is 30 frames long and the DSS (used for geometry capture) comprises 184 frames. In the second sequence, the test subject wears a diffuse t-shirt and slightly specular trousers, and performs a basic taichi motion, Fig. 1 and 5c. While the RES contains again 30 frames, the actual motion in the DSS is 110 frames long.

As can be seen in Fig. 5 and Fig. 1, and also in the accompanying video (supplementary material [1]), our reconstructed actor model faithfully captures even subtle detail, in particular wrinkles in clothing and folds, as *true* geometry. Fig. 4 zooms in on certain areas of the body model to illustrate that our MRF-based fusion method allows for reconstruction of subtle folds whose width is in the range of a few millimeters. This is a major improvement in shape quality over the original smooth template which was lacking any such detail, Fig. 1b and Fig. 4a,e. We would also like to point out that our final result is not only very detailed and almost free of artifacts at individual time steps, but due to the spatio-temporal MRF framework also faithful and smooth over time, see video [1]. The latter shows the unprecedented ability of our method to generate spatio-temporally smooth and detailed results even in the presence of measurement noise.

Although our visual results show qualitatively that we can measure highly-accurate scene geometry at sub-triangulation resolution, we also want to provide a more elaborate validation. Unfortunately, there exists no other scanning technology that would provide us with ground truth dynamic geometry at the same level of detail.

We therefore resort to another data set kindly provided to us by Theobalt et al. This data set contains an RES in which the actor strikes a static pose on a rotating turntable. In addition to the recording of the RES, a laser scan of the person

was taken during preprocessing. Since we were also given the pose of the template at each frame, we were able to reconstruct the BRDF and normal map based on our method, and could use our MRF framework to generate detailed surface shape. Since the scan and template possess different triangulations direct vertex comparison is infeasible. However, visual comparison of our result Fig. **Additional 1**e (see supplementary material [2]) and the scanned ground truth Fig. **Additional 1**b shows that all detail present in the original scan is also present in the deformed template, and that the resolution at which geometry was recovered is even higher in our result.

The detailed geometry we deliver is not only beneficial in high-quality animation applications, but also during 3D video rendering. Since our final geometry is much closer to the ground truth, it can be seen in Fig. **Additional 2**b that even simple projective texturing of our shapes produces better surface appearance than on the original template shown in Fig. **Additional 2**a.

Typically, we reconstruct as many as 350,000 displacement values over the template surface. Even at this detail level and when using a small patch size of 16 pixels, it takes moderate 5 to 6 minutes per time step of video to find the final detailed surface. Optimal values for the parameters α, β, γ and δ were found experimentally. To this end, we used a sequence of 3 of the reconstructed detailed meshes of sequence 1 as a ground truth and used their normal fields as measured normal fields. reconstruction errors could now be measured for a reasonable sample of combinations of the coefficients. Optimal results are obtained for $\alpha = 0.6$, $\beta = 0.1$, $\gamma = 0.2$ and $\delta = 0.1$ which were used in all our experiments.

Our method is subject to a couple of limitations. An important assumption enabling us to properly localize our final geometric solution in space is the one that the template is close to the true geometry. Unfortunately, this assumption is not entirely true for the head of the template as there may be quite some differences to true hair style and face geometry. Simple free-form deformation as performed by [19] cannot compensate for this. Therefore, we exclude the head from our reconstructions and note that this is a problem attributed to the provided input data.

Secondly, the current template employed by Theobalt et al. limits the types of scenes that we can handle to people wearing not too wide apparel. However, this is not a general limitation of our own contribution as we can easily apply our method to coarse geometry reconstructed with any other approach as well, as long as the geometry (triangulation) is coherent over time.

The original taichi input sequence also shows some jitter in the pose of the smooth template (slightly noticeable in the video result [1]), possibly due to tracking inaccuracies. We did not take any measures to compensate for this.

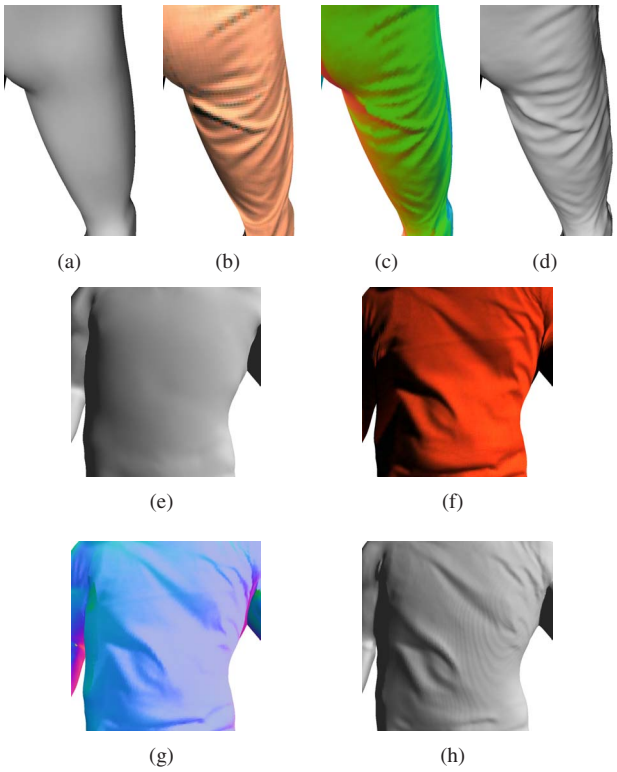


Figure 4. Our method can capture even subtle folds and wrinkles whose size is in the range of a few millimeters only. Zoom-in on leg: (a) smooth template, (b) template with texture, (c) color-coded normal field, (d) our final result rendered in OpenGL using Gouraud shading. (e)-(h) show a similar zoom onto the torso of the subject in the walking sequence. Also here, surface details were faithfully recovered in geometry.

Finally, in any frame where a surface point is in shadow from the light source, no normal direction can be reconstructed and the template normal is used instead. In the video [1], this effect is sometimes noticeable when the arm casts a shadow on the torso. However, our method handles this situation gracefully and produces the best possible result given this hard-to-prevent occasional lack of data in general moving scenes.

Despite these limitations we have presented one of the first approaches to reconstruct high-quality and high detail geometry of large dynamic scenes in a purely passive way.

6. Conclusion

We presented one of the first passive methods to reconstruct geometry of large dynamic scenes showing moving actors at unprecedented detail and accuracy from video only. To this end, in a first step we built on a previous method from the literature that allows for capturing of coarse geometry, surface reflectance and dynamic normal

(a) (b) (c)

Figure 5. Each pair of images shows, side-by-side, one original input video frame and the full 3D surface model with all geometric detail rendered in OpenGL from the same perspective. The direct comparison shows that our method captures even subtle dynamic geometric details in the actor's clothing very accurately.

maps. We then applied a new MRF-based spatio-temporal surface deformation approach that converts the geometric details encoded in the normals into true 3D displacements over the smooth template. Our method faithfully handles typical heavy-tail measurement noise, and is one of the first to allow for spatially accurate and temporally consistent height reconstruction over curved dynamic base geometry.

Appendix

The Huber function $\mathcal{H}$ is defined as

$$\mathcal{H}(R) = \begin{cases} \frac{1}{2}R^2 & \text{, if } |R| \leq k \\ k|R| - \frac{1}{2}k^2 & \text{, if } |R| > k \end{cases} \quad (9)$$

where k is the clip threshold [12]. $\frac{d\mathcal{H}}{d\mathcal{R}}$ is continuous and often referred to in the literature as the clip function.

References

[1] http://www.mpi-inf.mpg.de/∼nahmed/CVPR08b.wmv .

[2] http://www.mpi-inf.mpg.de/∼nahmed/CVPR08b-SM.pdf .

[3] A. K. Agrawal, R. Raskar, and R. Chellappa. What is the range of surface reconstructions from a gradient field? In *Proc. of ECCV*, pages 578–591, 2006.

[4] R. H. Byrd, P. Lu, J. Nocedal, and C. Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208, 1995.

[5] J. Y. Chang, K. M. Lee, and S. U. Lee. Multiview normal field integration using level set methods. In *CVPR*, 2007.

[6] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *CVIU*, 96(3):367–392, 2004.

[7] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. PAMI*, 10(4):439–451, 1988.

[8] P. Fua and Y. G. Leclerc. Using 3-dimensional meshes to combine image-based and geometry-based constraints. In *Proc. of ECCV*, pages 281–291, 1994.

[9] A. S. Georghiades. Recovering 3-d shape and reflectance from a small number of photographs. In *Proc. of EGSR*, pages 230–240, 2003.

[10] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying brdfs from photometric stereo. In *Proc. of ICCV*, pages 341–448, 2004.

[11] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *Proc. of ICCV*, 2007.

[12] P. Huber. *Robust Statistics*. Wiley, 2004.

[13] A. Jones, A. Gardne, M. Bolas, I. McDowall, and P. Debevec. Performance geometry capture for spatially varying relighting. In *Proc. of CVMP*, 2006.

[14] E. P. F. Lafortune, S.-C. Foo, K. E. Torrance, and D. P. Greenberg. Non-linear approximation of reflectance functions. In *Proc. of SIGGRAPH'97*, pages 117–126. ACM Press, 1997.

[15] H. Lange. Advances in the cooperation of shape from shading and stereo vision. *3dim*, 00:0046, 1999.

[16] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM TOG*, 22(2):27, 2003.

[17] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM TOG*, 24(3), 2005.

[18] J. Starck, G. Miller, and A. Hilton. Volumetric stereo with silhouette and feature constraints. *Proc. of BMVC*, 3:1189–1198, 2006.

[19] C. Theobalt, N. Ahmed, H. P. A. Lensch, M. Magnor, and H.-P. Seidel. Seeing people in different light - joint shape, motion and reflectance capture. *IEEE TVCG*, 2007.

[20] R. J. Woodham. Photometric method for determining surface orientation from multiple images. pages 513–531, 1989.

[21] R. Zayer, C. Rössl, and H.-P. Seidel. Discrete tensorial quasi-harmonic maps. In *Proc. of SMI*, pages 276–285, 2005.

[22] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *ACM TOG*, pages 548–558, 2004.

[23] R. Zhang, P.-S. Tsai, J. Cryer, and M. Shah. Shape from Shading: A Survey. *IEEE Trans. PAMI*, 21(8):690–706, 1999.