Full Body Performance Capture under Uncontrolled and Varying Illumination : A Shading-based Approach

Chenglei Wu[1,2], Kiran Varanasi[1], and Christian Theobalt[1]

[1]Max Planck Institute for Informatik
[2]Intel Visual Computing Institute

Abstract. This paper presents a marker-less method for full body human performance capture by analyzing shading information from a sequence of multi-view images, which are recorded under uncontrolled and changing lighting conditions. Both the articulated motion of the limbs and then the fine-scale surface detail are estimated in a temporally coherent manner. In a temporal framework, differential 3D human pose-changes from the previous time-step are expressed in terms of constraints on the visible image displacements derived from shading cues, estimated albedo and estimated scene illumination. The incident illumination at each frame are estimated jointly with pose, by assuming the Lambertian model of reflectance. The proposed method is independent of image silhouettes and training data, and is thus applicable in cases where background segmentation cannot be performed or a set of training poses is unavailable. We show results on challenging cases for pose-tracking such as changing backgrounds, occlusions and changing lighting conditions.

1 Introduction

Marker-less capture of human skeletal motion from images is one of the well-studied problems of computer vision, with recent advances being able to reconstruct human motion at increasing speed and accuracy and under lesser controlled situations [1–7]. These methods have several applications in industry: ranging from game and movie productions to use in biomechanics, ergonomy and sports sciences. However, despite great algorithmic advances, even latest approaches can not yet be applied in arbitrary environments with possibly changing lighting conditions, occlusions and starkly varying scene backgrounds. This is why purposefully placing markers in the scene is still the method of choice under such more challenging conditions [8]. Special effects professionals and producers of 3D video content are sometimes interested beyond kinematic motion parameters - demanding faithful and detailed dynamic 3D shape models of captured scenes, such that believable virtual actors or convincing novel viewpoint renderings can be created. The research community has responded to this requirement by developing so-called *performance capture* approaches, *i.e,* methods that simultaneously capture shape, motion and possibly appearance of people in

general apparel from a handful of video recordings [9–13]. Unfortunately, many state-of-the-art performance capture approaches are limited to studio settings with controlled lighting, controlled background, and to scenes without static or dynamic occluders. This has prevented the use of performance capture in practical applications such as outdoor movie sets or sports stadiums.

In this paper, we make a principal contribution towards the goal of model-based performance capture under less controlled conditions. We propose an algorithm that analyzes shading information to simultaneously estimate (a) human skeletal motion parameters, (b) arbitrary and time-varying incident scene illumination, (c) an approximation of surface reflectance, and (d) detailed dynamic shape geometry - such as folds and muscle bulges. We accept as input a multi-view video recorded from a synchronized and calibrated set of cameras, along with a rough initial shape-template of the person given as a 3D mesh fit to a kinematic skeleton. We do not require the subject to wear specific clothing or markers. Unlike previous performance capture methods [9–13], we do not require a fully controlled scene background, such as green screen, and thus do not expect exact foreground-background segmentations. We handle changing background and even some occlusions in the scene (Fig. 1). We do not rely on image features such as SIFT; our method is suitable even when the subject wears sparsely textured clothing.

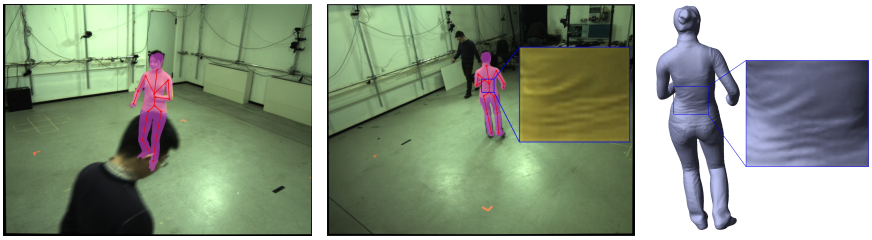
The main idea in our paper is to mathematically formulate the image shading constraint in terms of its differential towards the motion parameters of the kinematic chain representing human body pose. Along with pose, we simultaneously estimate time-varying incident illumination, surface albedo and detailed surface geometry in a joint framework. Thus, we integrate the human motion estimation problem into the broader framework of multi view shape-from-shading.

Our major contributions in this paper are as follows.

1. We present a new theoretical formulation of performance capture that simultaneously recovers human articulated motion and time-varying incident illumination, by a minimization of shading-based error.
2. We provide a solution to reconstruct both skeletal motion estimates and finely detailed time-varying 3D surface geometry for human performances that are recorded under general and changing illumination and in front of less constrained background.

2 Related work

For a thorough discussion and a historical perspective on human motion capture from images, one should consult any of the surveys [14, 7, 5]. Research efforts today can be broadly distinguished into studio-based methods which use multiple synchronized and calibrated cameras to achieve a high level of accuracy, and general purpose methods that work under fewer cameras in potentially cluttered surroundings - albeit producing pose estimates of lower accuracy. Many of the successful methods [15–17] validated on the HumanEva dataset [5] rely on a set



(a) (b) (c)

Fig. 1. Shading based pose tracking: (a,b) Overlay of estimated pose with recorded images - the actor is partially occluded by a person moving in the background (c) Reconstructed high-detail 3D geometry. The inset shows folds of the yellow T-shirt captured in 3D.

of training poses of tracking which limits their generalizability to new poses not observed in the training set. Methods for *performance capture* [9–13], *i.e,* detailed reconstruction of 3D surfaces along with skeletal motion, required studio conditions and green-screen to facilitate background segmentation. By contrast, in this paper we propose a shading-based approach that requires neither silhouettes nor training data. Silhouette estimation is sometimes integrated into the pose-estimation pipeline [18, 19] where 3D shape estimates are incorporated as a prior into image segmentation step. In particular, Hasler et al. [18] use this idea for an outdoor motion capture method. However, their approach does not capture detailed time-varying surface geometry. Also, background segmentation is an inherently error-prone step that fails in many cases; and hence should be avoided if possible for 3D shape reconstruction. Stoll et al [6] recently proposed a sums-of-Gaussian based holistic image and shape representation for pose tracking without silhouettes. But unlike them, we handle dynamic lighting changes, and recover not only body pose but also dense 3D surface detail, by analyzing image shading information.

Works in dynamic photometric stereo [20, 21] relied on specially engineered illumination to recover normal orientations that could be integrated to obtain the 3D surface. For example, a light-stage [21] captures images under temporally multiplexed illumination : with the shape being recorded under multiple known lighting conditions that provide a basis for describing light variations. These works analyze image shading information at a dense scale and thus recover true dynamic surface detail, instead of interpolating it from sparse image information such as silhouettes. However, finding a temporally coherent parameterization of the dynamic surface, despite some recent efforts [22, 23], remains a difficult task - especially when skeletal articulated motion need also be simultaneously captured. Wilson et al [24] use stereo and optical flow in a light-stage setup to obtain a temporally coherent parameterization for facial performance capture. They compute optical flow amidst a subset of *tracking frames* that are all captured under the same incident lighting. By contrast, in this work, we address

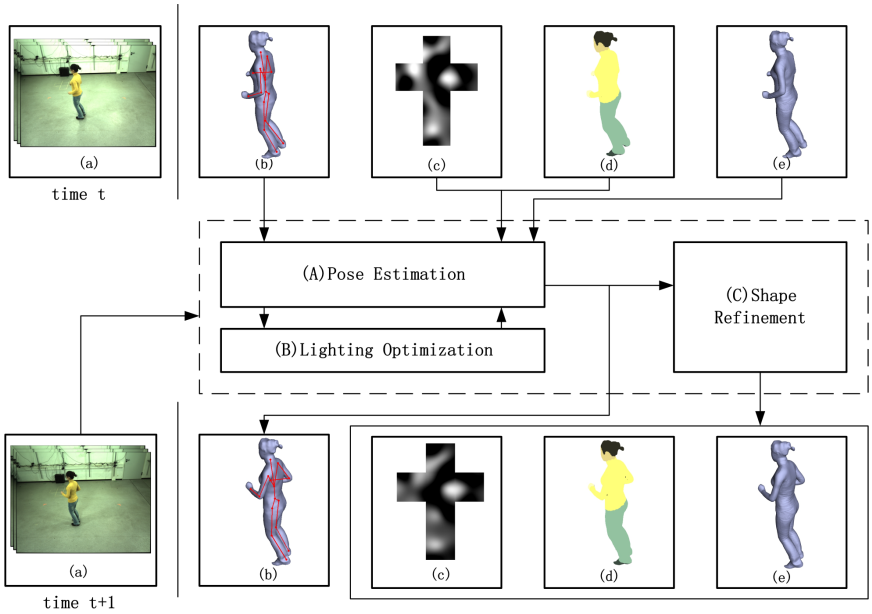


Fig. 2. Overview: (a) input multi-view images (b) skeletal pose (c) incident illumination (d) surface albedo (e) refined surface geometry. (b-e) are outputs of our method. Steps (A,B) for estimating pose and lighting are alternated in a joint optimization framework. In the step (C), final estimates of lighting, albedo and surface geometry are obtained. These estimates at t are provided as input for the optimization at $t+1$.

arbitrary and unknown lighting conditions which can vary from frame-to-frame. Wu et al.[25] have recently published a work that combines the strengths of model based performance capture with the inverse-rendering approaches of photometric stereo to reconstruct dynamic 3D surface detail that approaches the quality of light-stage reconstructions, albeit under arbitrary and unknown lighting conditions. The reconstructed surfaces are temporally coherent and aligned with simultaneous skeletal motion estimates. However, in that method, performance capture of the coarse geometry and dynamic shape refinement were treated as subsequent and independent problems, and the first part required the scene to be covered in green screen to enable coarse geometry estimation via silhouettes. By contrast, we use illumination estimation and shading constraints throughout the performance capture pipeline, *i.e,* for skeletal pose estimation and detailed shape reconstruction.

Our work is also relevant to the broader problem of dynamic shape from shading. Zhang et al.[26] provide an elegant formulation for shape and motion estimation under varying illumination, but the number of unknowns in the problem make it severely under-constrained, limiting their approach to only rigid motion estimates. However, as mentioned by them and others, shading variations provide cues for estimating flow even in texture-less regions. In this work, we

build upon this insight to estimate complete articulated human motion under unknown and time-varying incident illumination, without relying on silhouettes or training data. To the best of our knowledge, this has not been attempted before, to achieve results of even lower quality.

3 Overview

The input to our method is a multi-view video sequence of a moving actor captured using a sparse set of synchronized and calibrated cameras. Lighting in the scene can be arbitrary and time-varying, and since no background subtraction is required, no green-screen is expected and other potentially occluding elements can be in the scene. A rigged 3D mesh model with an embedded skeleton is provided as a template for tracking. We only need a smooth template mesh at a low resolution; the fine-scale detail is added later by our method. Similar to [13], the smooth template is built from a static laser scan of a person, alternatively image-based reconstruction methods are also feasible. The embedded bone skeleton as well as the skinning weights for each vertex, which connect the mesh to the skeleton, are obtained using standard tools.

An outline of the processing pipeline is given in Fig. 2. Given a set of captured multi-view images (a) as input, at each time-step $t+1$ we estimate skeletal pose (b), incident illumination (c), surface albedo (d), and detailed surface geometry (e). For each of these variables, we solve an inverse-rendering problem that attempts to make the rendered images as-close-as-possible to the captured image data. In Step-A, starting with the skeleton and the refined mesh from time t, the skeletal pose is optimized by assuming incident lighting and surface albedo from t, thereby exploiting temporal coherence. In Step-B, the incident illumination at time $t+1$ is estimated based on the skinned coarse mesh in the new skeletal pose. The Step-A is then repeated by taking the newly estimated lighting which results in a better pose estimate. The steps A and B constitute the main part of our method and are described in Sec. 5. In Step-C, we re-estimate incident lighting, surface albedo and then refine the surface geometry. The refined surface now captures folds and bulges not describable by articulated motion. For the initialization of the very first frame, we refer readers to Gall et al. [13] for pose estimation based on the manually segmented silhouettes and Wu et al. [25] to calculate the albedo value for each albedo segment, which could be provided by the user or any albedo segmentation method.

4 Image Formation Model

Assuming the object being tracked is a non-emitter of light (*i.e,* no surface inter-reflections), the reflectance equation describing the light transport at a certain surface point on the object can be defined as [27]

$$I(q,\omega_o) = \int_{\Omega} L(\omega_i) V(q,\omega_i) \rho(q,\omega_i,\omega_o) \max(\omega_i \cdot n(q), 0) d\omega_i, \tag{1}$$

where $I(q, \omega_o)$ is the reflected radiance, and the variables q, $\boldsymbol{n}$, ω_i and ω_o are the spatial location, the surface normal, and the incident and outgoing light directions, respectively. The symbol Ω represents the domain of all possible directions, $L(\omega_i)$ represents the incident lighting, $V(q, \omega_i)$ is a binary visibility function, and $\rho(q, \omega_i, \omega_o)$ is the bidirectional reflectance distribution function of the surface at q. To simplify the reflectance equation, we assume the reflectance to be Lambertian *i.e,* $\rho(q, \omega_i, \omega_o) = \rho(q)$, and represent the light transport with spherical harmonics (SH) so that the integral in the spatial domain will be converted to a dot product in the frequency domain.

We define the variable $G = LV$ and represent it with SH coefficients g_k. Then Eq. (1) will be simplified as follows:

$$I(q) = \rho(q) \sum_{k=1}^{d^2} g_k(q) S_k(n(q)), \tag{2}$$

where $S_k(n(q))$ is the scaled SH basis function depending on the surface normal directions $n(q)$, and $d-1$ is the order of SH used. When visible lighting and albedo are known, the rendering value is determined by the surface normal only. This equation is employed to provide the shading constraints for pose estimation (Sec. 5) and later used for surface geometric refinement (Sec. 6).

5 Pose Estimation Under Varying Illumination

At each time-step $t+1$, we perform a simultaneous estimation of body pose and incident lighting, both of which may change from time t. In order to keep the optimization tractable, we assume that changes in body pose are independent from changes in lighting, and alternate between the optimization of these variables.

We take as initialization the refined mesh and the embedded skeleton of time t, as well as the estimated incident lighting and surface albedo. In Sec. 5.1, we introduce how the mesh changes according to pose-changes. In Sec. 5.2, we define the shading constraint used to estimate the pose parameters, given the incident lighting. The optimization to minimize the shading error is described afterwards. The method to estimate incident lighting is described in Sec. 5.3.

5.1 Surface parameterization with respect to pose

We use the popular linear blend skinning approach to deform the mesh to a skeletal pose. Similar to [1], we represent the articulated pose to be estimated by a set of twists $\theta_k \hat{\xi_k}$. The state of a kinematic chain is determined by a global twist $\hat{\xi}$ and the joint angles $\Theta = (\theta_1, \cdots, \theta_m)$. Assuming the state of the kinematic skeleton of the previous time-step to be known, the unknowns for pose estimation are the rigid motion of the root node and changes in joint angles which we denote as $\phi = (\Delta\hat{\xi}, \Delta\theta_1, \cdots, \Delta\theta_m)$. Let q_i^t be the position of vertex i at t. By using exponential maps to represent each joint's rigid motion and by linearizing the

rigid body transforms, the pose of the vertex i at $t+1$ can be expressed with the skinning equation as

$$\begin{aligned} \begin{pmatrix} q_i^{t+1} \\ 1 \end{pmatrix} &= \sum_{j=1}^{m} w_j e^{\Delta\hat{\xi}} \prod_{k \in T(j)} e^{\hat{\xi_k} \cdot \Delta\theta_k} \begin{pmatrix} q_i^t \\ 1 \end{pmatrix} \\ \approx \begin{pmatrix} q_i^t \\ 1 \end{pmatrix} + \left(\Delta\hat{\xi} + \sum_{j=1}^{m} w_j \sum_{k \in T(j)} \hat{\xi_k} \cdot \Delta\theta_k \right) \begin{pmatrix} q_i^t \\ 1 \end{pmatrix} &= \begin{pmatrix} q_i^t \\ 1 \end{pmatrix} + M_q(i) \cdot \phi, \end{aligned} \tag{3}$$

where $T(j)$ determines the indices of joints preceding the joint k in the kinematic chain, and $M_q(i)$ is the matrix determining how the pose change influences the change of vertex position. Each vertex i is assigned a set of skinning weights w_j that determine how much influence bone (or joint) j has on the deformation of vertex i. Skinning weights are once defined during template building using standard techniques [13]. A similar equation can be derived for the vertex normal n_i^{t+1} at time $t+1$

$$\begin{pmatrix} n_i^{t+1} \\ 0 \end{pmatrix} \approx \begin{pmatrix} n_i^t \\ 0 \end{pmatrix} + M_n(i) \cdot \phi, \tag{4}$$

where $M_n(i)$ is a matrix that determines how the pose change ϕ results in a change in normal orientation.

5.2 Shading constraint for pose estimation

Our shading constraint requires the rendered images of the optimal pose according to our lighting model to be as-close-as-possible to the image data captured. Following Eq. (2), the shading constraint for a single camera c is defined as

$$E_c^s = \sum_i (\rho_i g(q_i^{t+1}) \cdot S(n_i^{t+1}) - I_c^{t+1}(x_i^{t+1}, y_i^{t+1}))^2, \tag{5}$$

where (x_i^{t+1}, y_i^{t+1}) is the projection of the surface vertex q_i^{t+1}, and $g(q_i^{t+1})$ and $S(n_i^{t+1})$ are the vectors of SH coefficients g_k and S_k of Eq. (2). We assume the albedo ρ_i at time $t+1$ is the same as that at time t, thereby exploiting temporal coherence in scene motion. However, both the lighting and geometry at time $t+1$ are unknown. We attempt to estimate both of them in a unified framework in order to properly account for shading changes due to changes in either lighting or pose. Since simultaneous estimation of both of them is computationally challenging, we alternate between error minimization with respect to either of these two variables. First we minimize the shading error to estimate the pose, by assuming the lighting of the previous time-step, and thereafter we solve for lighting. To do this, we linearize the SH term $S(n_i^{t+1})$ and the image intensity term I_c^{t+1}. The SH term is expressed in a first-order Taylor-series expansion, and using the terms of Eq. (4).

$$S(n_i^{t+1}) \approx S(n_i^t) + \frac{\partial S(n_i^t)}{\partial n_i^t} \Delta n_i^t = S(n_i^t) + \frac{\partial S(n_i^t)}{\partial n_i^t} M_n(i) \cdot \boldsymbol{\phi}, \tag{6}$$

where $\frac{\partial S(n_i^t)}{\partial n_i^t}$ is derivative of scaled SH function with respect to normal changes Δn_i^t, which are expressed in terms of pose changes ϕ.

Inspired by the formulation of optical flow, we linearize $I^{t+1}(x_i^{t+1}, y_i^{t+1})$ as:

$$I^{t+1}(x_i^{t+1}, y_i^{t+1}) = I^{t+1}(x_i^t + u_i, y_i^t + v_i) \approx I^{t+1}(x_i^t, y_i^t) + I_x^{t+1} u_i + I_y^{t+1} v_i. \quad (7)$$

Next, we derive the linear approximation for the flow (u_i, v_i) in an image from the motion parameters ϕ. This is similar to the derivation in [1], but we use the full perspective camera model instead of scaled orthographic projection [1], as camera calibration is available for our system. Then, the image motion from time t to time $t+1$ can be linearized as:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \approx \begin{pmatrix} \frac{s_1}{Z_i^t} & 0 & 0 & s_3 \\ 0 & \frac{s_2}{Z_i^t} & 0 & s_4 \end{pmatrix} \cdot \boldsymbol{e}^{\hat{\xi_c}} \cdot \begin{pmatrix} \Delta q_i^t \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{s_1}{Z_i^{t2}} & 0 & 0 & 0 \\ 0 & \frac{s_2}{Z_i^{t2}} & 0 & 0 \end{pmatrix} \cdot \boldsymbol{e}^{\hat{\xi_c}} \cdot \begin{pmatrix} q_i^t \\ 1 \end{pmatrix} \cdot \Delta Z_i^t, \quad (8)$$

where s_1, s_2, s_3, s_4 are the acquired camera intrinsic parameters, $\boldsymbol{e}^{\hat{\xi_c}}$ acts as the extrinsic matrix of the camera's pose, Z_i^t is the depth of q_i^t for the current camera. The linearization is based on the assumption that the rigid motion Δq_i^t as well as the relative depth change ΔZ_i^t are small enough. As both of them can be expressed through pose change ϕ (from Eq. (3)), the flow (u_i, v_i) can ultimately be expressed as a linear function of ϕ.

The shading constraint in Eq. (5) can be further improved by considering the color similarity between the rendered color and the image color. The color similarity is computed as the Euclidean distance in HSV space and appears as a weighting factor α_i in our shading constraint. This helps us avoid optimizing the model where the template material does not yet match to its projection in the image. Combining terms from multiple cameras, our non-linear multi-view shading energy function is given as

$$E = \frac{1}{N} \sum_c \sum_i \{\alpha_i^c (\rho_i g(q_i^{t+1}) \cdot S(n_i^{t+1}) - I_c^{t+1}(x_i^{t+1}, y_i^{t+1}))\}^2, \quad (9)$$

where N is the total number of constraints for error normalization (*i.e,* the number of pixels in all cameras getting the projection from the mesh), and α_i^c is the color similarity for pixel i in camera c. Using the previously described recipe of linearization, this can be expressed in terms of pose parameters ϕ as a linear system:

$$\boldsymbol{H} \cdot \phi = \boldsymbol{b} \quad (10)$$

Specifically, the k^{th} rows of matrix $\boldsymbol{H}$ and vector $\boldsymbol{b}$ have the following form (detailed derivation is in the supplementary document, $r_3^\top$ refers to the last row of the rotation matrix of the camera pose) :

$$\begin{aligned} \boldsymbol{H}_k = \alpha_i^c \rho_i g(q_i^{t+1}) \cdot \frac{\partial S(n_i^t)}{\partial n_i^t} M_n(i) - \alpha_i^c \left[\tfrac{s_1}{Z_i^t} I_x^{t+1},\ \tfrac{s_2}{Z_i^t} I_y^{t+1},\ 0,\ s_3 I_x^{t+1} + s_4 I_y^{t+1} \right] \boldsymbol{e}^{\hat{\xi_c}} M_q(i) \\ + \alpha_i^c \left[\tfrac{s_1}{Z_i^{t2}} I_x^{t+1},\ \tfrac{s_1}{Z_i^{t2}} I_y^{t+1},\ 0,\ 0 \right] \boldsymbol{e}^{\hat{\xi_c}} \begin{bmatrix} q_i^t \\ 1 \end{bmatrix} \cdot \begin{bmatrix} r_3^T & 0 \end{bmatrix} \cdot M_q(i), \\ \boldsymbol{b}_k = \alpha_i^c I^{t+1}(x_i^t, y_i^t) - \alpha_i^c \rho_i g(q_i^{t+1}) \cdot S(n_i^t). \end{aligned} \quad (11)$$

Coarse-to-Fine Optimization To minimize the non-linear error function of Eq. (9), we iteratively solve Eq. (10) and linearize around the new solution. Note that here after solving Eq. (10), we check if the original energy in Eq. (9) decreases to decide the appropriate step size for updating the solution, in a fashion similar to Newton-Raphson style minimization with adaptive step size. Besides, as given in Eq. (7), the linearization assumes that the local image intensity variations can be approximated by a first-order Taylor expansion. So we adopt a coarse-to-fine strategy for pose estimation - by building an image pyramid through successively downsampling each captured image, and running the pose estimation from coarsest images to the finest images. This helps us track big motions and reduces the chance of getting stuck in local minima.

5.3 Lighting optimization

In the general case, lighting changes can be abrupt and impossible to model. However, for most cases, it can be assumed that the lighting at $t+1$ changes gradually from lighting at t. In our method, we optimize for pose and lighting in a two pass strategy. For the first pass, we use the lighting at t to optimize for pose at $t+1$, as described in the previous section. For the second pass, we estimate the lighting at $t+1$ based on the new pose, and then use it to refine the pose estimates. We have empirically observed that one additional iteration of alternating optimization is sufficient for getting good estimates.

We derive the constraint for lighting optimization from the image formation model defined in Eq. (1). But instead of Eq. (2), following Wu et al. [25], we use a different type of linearization. We define $T(q,\omega_i) = V(q,\omega_i)\max(\omega_i \cdot n(q), 0)$ and then represent it with SH coefficients t_k, while representing the incident lighting L with SH coefficients l_k. This gives the linearization:

$$I(q) = \rho(q) \sum_{k=1}^{d^2} l_k t_k. \tag{12}$$

We compare the rendered intensity values with the captured image I_c and solve for the lighting coefficients l_k. In order to deal with outliers, i.e. erroneous projection due to the inaccuracy of the pose, we solve a ℓ_1 norm minimization problem defined as:

$$\hat{l} = \underset{l}{\operatorname{argmin}} \sum_i \sum_{c \in Q(i)} |\sum_{k=1}^{d^2} l_k t_k - I_c(P_c(x_i))|. \tag{13}$$

Here, i is the vertex index, c is the camera index, $Q(i)$ is the set of cameras that can see the i-th vertex $\boldsymbol{x}_i$, and P_c is the projection matrix for camera c.

6 Dynamic Surface Refinement

After the pose and lighting estimation step, we have a coarse template model that strikes the correct pose, as parameterized by the respective skeleton pose

parameters. Different from linear skinning that we used in skeletal pose estimation for its simplicity, we here use quaternion blend skinning[28] to render the final shape of the surface mesh in the current pose, as it leads to higher quality surface deformation, in particular around joints. When we have the coarse mesh of time $t+1$, we refine the vertex positions q_i from shading cues as given in Eq. (3). We refer to [25] for detailed explanation of this step. A minor difference is that temporal consistency is taken into account for assigning albedo labels, by formulating this as a Markov-Random-Field (MRF) problem with the data term consisting of two values (i) the similarity of vertex color to the average color in the material label and (ii) the label similarity with previous time-step.

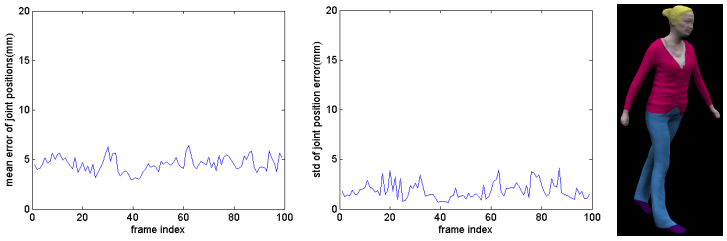
7 Results

7.1 Quantitative Evaluation

In order to quantitatively evaluate our method, we generated a synthetic sequence of 100 frames with 10 camera views. The ground-truth skeleton and mesh geometry are taken from the results of a previous performance capture method of a human walking sequence. The ground-truth surface albedo map and dynamically changing illumination are manually assigned. With these generated synthetic images as input, and given the mesh, skeleton, albedo segmentations for the first frame, we run our algorithm on the remaining 99 frames. In Fig. 3, we report the accuracy of our approach, with the mean joint position error of only around 6 mm.

7.2 Real-word sequences

We use three real captured sequences for qualitatively evaluating our method. The sequences were captured with 11 cameras in a studio, but unlike in the input data of previous performance capture methods, the subject can wear sparsely textured apparel, there is no need for green-screen background, and there may be potentially occluding objects in the scene and dynamic background (Fig. 1). Cameras recorded at a resolution of 1296×972 pixels, and at a frame rate of $40fps$. Each sequence shows major illumination changes; they are induced by an operator randomly setting control knobs for various lights in the studio - these readings are not taken nor provided in any way to our method. Please also note that some of the captured images are saturated, which our method handles robustly. As can be seen in the overlayed images of our estimated skeleton and 3D shape in Fig. 4, good pose estimates are obtained despite the challenging scene conditions. Even when a few cameras are partially occluded, our method still works quite well thanks to the use of shading cues and multiple cameras setup. High quality surface detail such as deforming cloth folds are also captured (Fig. 6). We invite the readers to see the results in our accompanying video, which is better suited for observing temporal information. Minor errors in skeletal joint positions might cause the surface to jitter over time, which we remove in our video results by temporal smoothing of the vertices.



(a) (b) (c)

Fig. 3. Quantitative evaluation: (a) The mean error of joint positions. (b) The standard deviation of joint position errors. (c) A generated synthetic image.

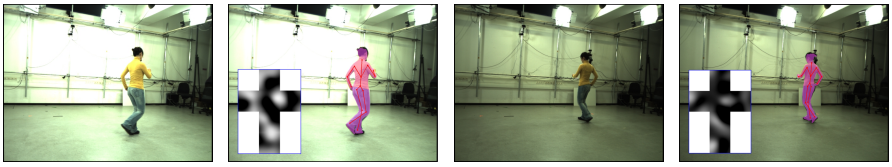
We compare the results of our method with a texture-based tracker that does not estimate lighting explicitly at each frame. Instead, it assumes texture from the first frame and uses optical flow for tracking; it loses track after a few frames as the lighting changes significantly (see Fig. 5-b). We also implemented a silhouette-based tracker [13] that explicitly performs background segmentation using chroma-keying on the captured images. Due to changing lighting and moving background objects, the extracted silhouettes are sometimes misleading and result in inaccurate pose estimates (see Fig. 5-c).

7.3 Computation Time

The computation time of our method depends on image resolution, mesh resolution and the order of SH used for representation. In our experiments, we represented 3D shape using meshes of 80000 vertices, and used a 4th order SH for representing lighting. With these values, our method takes about 10 min per each frame on a standard CPU with a 2.6 GHz processor and 8 GB RAM. Specifically, the computation times are 3 min for one-pass of pose estimation, which we do twice for each frame. The lighting estimation step is quite fast, taking only 10 seconds. The other time-consuming part is the dynamic shape refinement, which takes 4 min, of which 1 min is spent on visibility calculation. Striking a trade-off between representation accuracy and computation time, we utilized a low-resolution mesh (around 5000 vertices) to render the visibility map for each vertex on the high-resolution mesh. As our code is unoptimized, we believe the computational time can be further reduced by parallelizing the algorithm.

7.4 Limitations and Future Work

Our algorithm becomes less effective when the underlying shape template is not accurate. For example, the rotation of the upper arm may not be modeled in the skeleton. We corrected for such errors by manually adjusting the pose where the algorithm failed (roughly one frame per 200 frames needed such correction in our experiments). Please note that a global optimization strategy such as that



(a) (b) (c) (d)

Fig. 4. Illumination changes in a real captured sequence: (a,c) Frames showing widely different incident illumination (b,d) The output skeletal pose and mesh overlayed onto the images. The insets show estimated illumination at each frame.

(a) (b) (c)

Fig. 5. Comparison with alternative tracking methods: (a) Our method (b) Texture-based tracking (c) Silhouette-based tracking [13]

used in [13] can automatically handle such cases. Also since we estimate lighting and pose sequentially at each time-step, error accumulation may cause drift of the tracker. In future work, we would like to address this issue by stronger priors from data-driven modeling. Our assumptions of Lambertian reflectance and local shading model may not be justified in some cases. Abrupt lighting changes, e.g, the illumination generated by a controlled light stage, are also hard to model. However, in such cases, the lighting pattern is known beforehand and can be directly provided as input to our method. A final limitation is the computation time for running our method which is too high for real-time deployment. We would like to address these and other limitations in future work.

8 Conclusion

In this paper, we provide a novel shading based frame-work for human performance capture under uncontrolled and dynamic lighting. Starting from synchronized multi-view images, we estimate both the articulated human pose and fine-scale time varying surface geometry. Key innovation is a novel iterative pose optimization framework that exploits estimated lighting and shading cues. Our approach does not expect carefully engineered backgrounds as it does not perform silhouette extraction or any other form of background segmentation. Ultimately, one of the goals of vision based motion capture is to obtain high quality motion reconstructions using a very limited set of cameras in outdoor

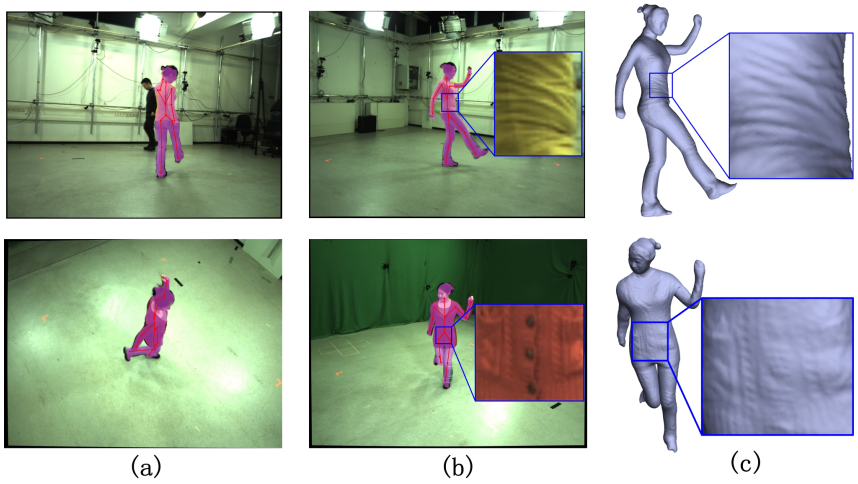


Fig. 6. Results of pose and 3D shape estimation: (a,b) Overlayed skeletal pose at different frames and camera views (c) Fine-scale 3D shape reconstruction. The inset shows dynamic cloth deformations captured from shading.

situations. Even though we do not explicitly evaluate our method in outdoor scenes, we believe that our work provides a crucial step towards this goal.

References

1. Bregler, C., Malik, J., Pullen, K.: Twist based acquisition and tracking of animal and human kinematics. IJCV **56**(3) (2004) 179–194
2. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In: ECCV. (2000) 702–718
3. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR. (2000) 1144–1149
4. Balan, A., Sigal, L., Black, M., Davis, J., Haussecker, H.: Detailed human shape and pose from images. In: CVPR. (2007)
5. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV **87** (2010) 4–27
6. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: ICCV. (2011)
7. Poppe, R.: Vision-based human motion analysis: An overview. CVIU **108**(1-2) (2007)
8. Raskar, R., Nii, H., deDecker, B., Hashimoto, Y., Summet, J., Moore, D., Zhao, Y., Westhues, J., Dietz, P., Barnwell, J., Nayar, S., Inami, M., Bekaert, P., Noland, M., Branzoi, V., Bruns, E.: Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. ACM Trans. Graph. **26** (July 2007)
9. Vlasic, D., Baran, I., Matusik, W., Popovic, J.: Articulated mesh animation from multi-view silhouettes. In: ACM TOG (Proc. SIGGRAPH). (2008) 97:1–97:9

10. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: Proc. SIGGRAPH. (2008)
11. Cagniart, C., Boyer, E., Ilic, S.: Free-form mesh tracking : a patch-based approach. In: CVPR. (2010)
12. Starck, J., Hilton, A.: Surface capture for performance based animation. IEEE Computer Graphics and Applications **27(3)** (2007) 21–31
13. Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton and surface estimation. In: CVPR. (2009)
14. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU **104**(2) (2006) 90–126
15. Li, R., Tian, T.P., Sclaroff, S., Yang, M.H.: 3d human motion tracking with a coordinated mixture of factor analyzers. IJCV **87** (2010) 170–190
16. Lee, C.S., Elgammal, A.: Coupled visual and kinematic manifold models for tracking. IJCV **87** (2010) 118–139
17. Bo, L., Sminchisescu, C.: Twin gaussian processes for structured prediction. IJCV **87** (2010) 28–52
18. Hasler, N., Rosenhahn, B., Thormählen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: CVPR. (2009)
19. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: CVPR. (2011)
20. Hernandez, C., Vogiatzis, G., Cipolla, R.: Multiview photometric stereo. IEEE TPAMI **30**(3) (2008) 548–554
21. Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T., Debevec, P.: Performance relighting and reflectance transformation with time-multiplexed illumination. ACM TOG (Proc. SIGGRAPH) **24**(3) (July 2005) 756–764
22. Vlasic, D., Peers, P., Baran, I., Debevec, P., Popovic, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. ACM TOG (Proc. SIGGRAPH) **28**(5) (2009) 174
23. Popa, T., South-Dickinson, I., Bradley, D., Sheffer, A., Heidrich, W.: Globally consistent space-time reconstruction. In: SGP. (2010)
24. Wilson, C., Ghosh, A., Peers, P., Chiang, J.Y., Busch, J., Debevec, P.: Temporal upsampling of performance geometry using photometric alignment. ACM TOG (Proc. SIGGRAPH) **29**(2) (March 2010)
25. Wu, C., Varanasi, K., Liu, Y., Seidel, H.P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In: ICCV. (2011)
26. Zhang, L., Curless, B., Hertzmann, A., Seitz, S.: Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. In: ICCV. (oct. 2003) 618 –625 vol.1
27. Kajiya, J.T.: The rendering equation. Proc. ACM SIGGRAPH **20**(4) (1986)
28. Kavan, L., Collins, S., Žára, J., O'Sullivan, C.: Skinning with dual quaternions. In: Symposium on Interactive 3D graphics and games. (2007) 39–46