Consistent 3D Human Shape from Repeatable Action

Keisuke Shibata† Sangeun Lee† Shohei Nobuhara†‡ Ko Nishino†
† Kyoto University ‡ JST PRESTO
https://vision.ist.i.kyoto-u.ac.jp/

(a) Video Capture (b) Video of Repeated Action (c) Reconstruction Results

Figure 1: We derive a novel method for reconstructing a clothed human body shape that is consistent across frames from videos capturing a repeatable action from a few viewpoints (a,b). We show the use of reconstructed human body shape for free-viewpoint rendering (c).

Abstract

We introduce a novel method for reconstructing the 3D human body from a video of a person in action. Our method recovers a single clothed body model that can explain all frames in the input. Our method builds on two key ideas: exploit the repeatability of human action and use the human body for camera calibration and anchoring. The input is a set of image sequences captured with a single camera at different viewpoints but of different instances of a repeatable action (e.g., *batting). Detected 2D joints are used to calibrate the videos in space and time. The sparse viewpoints of the input videos are significantly increased by bone-anchored transformations into rest-pose. These virtually expanded calibrated camera views let us reconstruct surface points and free-form deform a mesh model to extract the frame-consistent personalized clothed body surface. In other words, we show how a casually taken video sequence can be converted into a calibrated dense multiview image set from which the 3D clothed body surface can be geometrically measured. We introduce two new datasets to validate the effectiveness of our method quantitatively and qualitatively and demonstrate free-viewpoint video playback.*

1. Introduction

Image-based 3D human body reconstruction has a wide variety of applications. For instance, it can enable personalization of avatars and product designs, health and fitness monitoring, and visual media content creation such as free-viewpoint replay. Deep methods, in particular, have demonstrated remarkable progress by harnessing the power of learned priors of the human body structure both in its shape and articulation. Recent works (e.g., [43]) have shown that detailed clothed body shape including its unseen side can be recovered from a single image.

Recovering a person in action from a video is, however, not as trivial as applying these single-image methods to each frame as it would result in frame-varying, inconsistent reconstructions. This shape consistency, i.e., that we obtain a single shape that explains pose-varying body shapes in all frames, is critical for video-based 3D human shape reconstruction. Without such a consistent 3D human model, modified geometry (e.g., pose) in one frame will not match that in another frame let alone what the person would actually look like.

Consistent 3D human body shape reconstruction from video is, however, challenging, as it fundamentally requires simultaneous reconstruction of shape and action. These

two need to be decoupled such that we arrive at a single 3D shape model that explains all articulations of it in all frames of the video. Kanazawa et al. [24] take a principled step in this direction by using a statistical shape model (SMPL [31]) and an adversarial posed-body loss to tame the wide variability of human body shape and articulation with learned priors. The results are convincing, but the recovered human bodies are inherently naked. Recovering clothed human body shape with the same approach, however, is nontrivial as it would entail learning a statistical human clothed body model whose variation would unlikely have strong enough structural regularities for even a complex deep network to learn.

In this paper, we tackle the challenging problem of recovering a consistent 3D clothed human body model from a casually taken video of a person in action. We ask, instead of learning structural variations that seems infeasible for clothed shapes, can we actually "measure" the clothed body shape? Just from a short video of a person in temporally changing poses, can we recover one detailed clothed surface of the person such that it can be posed into every frame?

We realize this by exploiting the repeatability of actions. Many actions are repeatable. For instance, sports actions such as golf swings and baseball pitching can be repeated with more or less the same body movements. Even if the body action may not be repeatable as a whole, its atomic parts such as gesticulation are often repeatable. In fact, the repetition of body movements has been exploited for 3D pose estimation in the past [4,30,41]. In contrast, our goal is to simultaneously recover the dense 3D human body shape in clothes as well as its frame-by-frame articulation.

The input to the method is a set of videos each capturing a different instance of the repeated action from a distinct viewpoint. Such input data can be captured with a single camera, for instance, by having a friend capture one swing at a time as she moves around you while you repeat your batting swing. The video set collectively provides multiview data albeit an uncalibrated one. Each sequence is from a different viewpoint but of a different action instance that can have spatio-temporal variations in its execution.

Our key idea is to leverage the human body itself to calibrate this casually captured repeatable action video. We use the joints and bones of the target person to spatio-temporally localize the cameras and virtually transform their frame instances into a dense calibrated multiview setting. We first detect 2D joints in each frame and recover the 3D skeleton of the target while calibrating the cameras spatially (*i.e*., extrinsic camera calibration) and temporally (*i.e*., temporal alignment of videos across viewpoints). Next, for each bone of the 3D skeleton model, the camera location of each frame of each sequence is associated with the bone and rotated into the rest pose (*i.e*., T-pose) bone orientation. Finally, we use geometric contour intersection to robustly recover the 3D body shape, i.e., geometrically measure body surface points. To convert these 3D intersections into a dense body surface, we free-form deform a generic 3D body model. Hands, feet, and face are out of the scope of this work and are simply replaced with generic shapes.

We demonstrate the effectiveness of our method quantitatively on synthetic data and qualitatively on real data and show comparisons with per-frame learning-based reconstruction as well as related video-based methods. The results show that the method successfully reconstructs a pose-consistent 3D surface model faithful to the clothing and body shape of each target. We also demonstrate its application to free-viewpoint video that enables better examination of, for instance, a sports action, which directly showcases practical use of the method. We believe our method, particularly its ease of capture, opens new avenues of usage and would find applications in a variety of domains including entertainment, communication, and health.

2. Related Work

Multiview Reconstruction Starting from the pioneering work by Kanade *et al*. [23], many studies have been proposed for 3D human shape reconstruction from multiview images. Inspired by early studies on representative reconstruction cues such as photoconsistency [23] and silhouette constraints [8,9,33], most past methods combine these cues to leverage their complementary advantages, *e.g*., accurate reconstruction by photoconsistency and robust initialization by silhouette constraints [11, 21, 29, 45, 49]. While the majority of such approaches are tailored to indoor environments, Mustafa *et al*. [34] proposed an outdoor capture pipeline using synchronized and calibrated multiview cameras. These methods realize frame-wise 3D reconstruction of arbitrary body shapes. Because of their bottom-up and data-driven nature, they can reconstruct humans with additional items [23] and humans wearing complex clothing [45]. In these approaches, up to 100 synchronized cameras are used [45].

Single-Image Reconstruction Since estimating the 3D shape from its single 2D projection is an ill-posed problem, single-image reconstruction methods rely on prior knowledge of the target 3D shape. The use of a dedicated 3D shape model, *i.e*., a 3D scan of the target itself, is a simple but effective means to build such priors [13,48]. These methods deform a 3D shape of the target such that photoconsistency and silhouette constraints from sparse multiview cameras are satisfied.

Owing to the proliferation of 3D scanners and RGB-D sensors, various 3D human shape datasets and statistical 3D human models have been introduced [1,3,7,14,31,37]. These statistical human 3D models allow single-image 3D

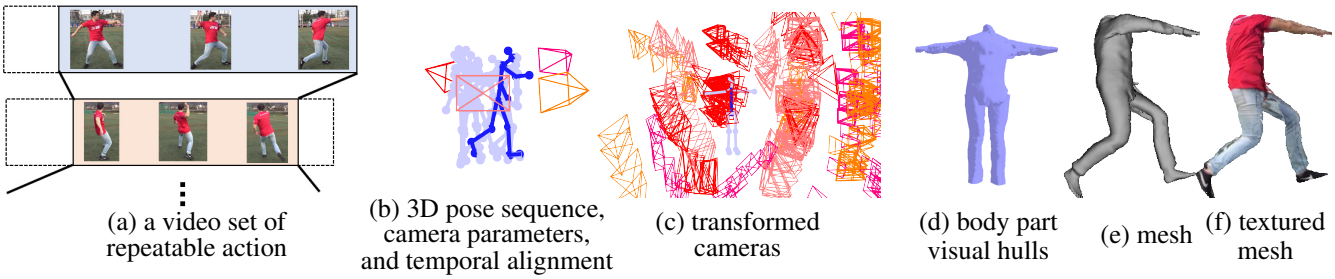

Figure 2. Overview of our approach. From a set of videos each capturing a different instance of a repeated action from a distinct viewpoint (a), we simultaneously recover human pose sequences, camera parameters, and temporal alignment (b). The 3D bones serve as anchors to the (bone-relative) viewpoint of each frame in each of the videos of different camera locations (each cone in distinct color, respectively) (c). These bone-anchored relative cameras are transformed into a common coordinate frame, which is used to construct a visual hull for each body part (d). A generic body surface model is free-form deformed to fit these visual hulls, resulting in a consistent clothed 3D body shape mesh model of the target person that can be rendered from a novel viewpoint with and without textures (e,f).

estimation, including silhouette-based [12, 18] and joint-based [6] approaches. Recent advances in deep learning also enabled single-image 3D shape estimation methods [19, 22, 24, 26, 27, 36, 42, 43, 47, 51] that implicitly encode the statistical knowledge on the 3D human shape and pose in neural networks. For example, Xiang *et al*. [51] introduced a method for monocular 3D human shape estimation including face and hands, and Saito *et al*. [42, 43] introduced a single-image clothed body estimation method and its extension to multiview inputs. Garau *et al*. [17] utilized this single-image approach for extrinsic camera calibration by estimating the relative posture between 3D shapes reconstructed at each view capturing the same target synchronously.

Repeating Motion As well known for static shape reconstruction, a moving camera orbiting around the target or a stationary camera capturing the target rotating in front of it can also provide multiview data for stereo or silhouette-based 3D reconstruction [2, 38, 44, 50]. Alldieck *et al*. [2] introduced 3D human body shape estimation from a single-view video by capturing a person in a static pose rotating in front of a camera. The method, however, cannot be applied to dynamic objects as the target would change its shape and pose.

If the target repeats a periodic motion, *e.g*., walking, while being captured, for each frame in a period, we can find corresponding frames in different periods that capture the same 3D shape from different viewpoints effectively. Belongie and Wills used this temporal periodicity for triangulating 3D human joints [4]. Ribnick and Papanikolopoulos reconstructed 3D trajectory of points of interest by exploiting the periodicity in 3D [41]. Li *et al*. proposed an algorithm that handles moving camera calibration and a-periodic target motions [30]. Dong *et al*. synchronize videos of repeated action using 3D pose estimates from a single camera, and improves the 3D poses with iterative optimization [15]. We show that dense clothed body surface can be recovered from the repetition of an action as a whole.

3. Repeatable Action

One key idea of our method is to exploit the repeatability of human actions. Many actions performed by people, especially those that may benefit from close examination afterwards, can be repeated with more or less spatially and temporally similar body movements. Examples include golf swings, skateboard tricks, baseball pitches, and soccer shots for sports, greeting a person, opening and entering a door, rising from bed, and sitting down on a chair for daily actions. We call these repeatable actions.

If we capture a repeated repeatable action one instance at a time from a fixed viewpoint, but from a distinct one for each instance, we already have sparse multivew data. We typically use 4 viewpoints. This multivew data is, however, very sparse and uncalibrated both in terms of the camera locations as well as the actual action of the person as each repeatable action instance is not exactly the same. We later show that this sparse multivew uncalibrated data can be turned into a dense calibrated multivew dataset. If we capture K instances of a repeatable action each with, on average, N frames of video, we show that we can obtain $K \times N$ views for each part of the body. That is, we show how the actual sparse views can be multiplied by tens coming from the number of frames in each video (e.g., 4 views turned into 360 views with 3 seconds 30fps videos).

Capturing such repeatable action is easy and can be done in a casual setting as it does not require synchronized simultaneous image capture. It can be done with a single camera moved to different vantage points around a person for each instance. That can be done by a friend with a phone camera or even alone with access to a tripod.

We make only two mild assumptions for capturing re-

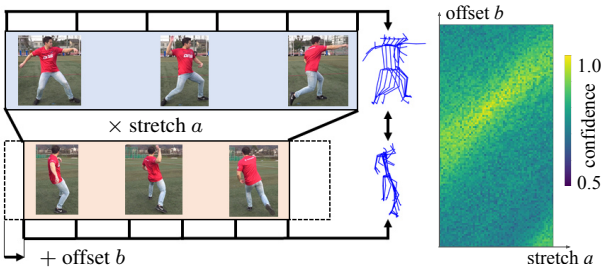

Figure 3. (a) We temporally calibrate repeatable action videos by evaluating and searching for the maximum inlier number of Sampson Distance [20] between pose sequences, which is used as the confidence of hypothesized temporal linear stretch and offset, and essential matrix (Eq. 1). (b) These confidence values show a clear global optimum corresponding to the correct spatio-temporal calibration of the videos.

peatable actions. First, we assume that the instances of a repeatable action vary in temporal duration but not in their local speeds. That is, we assume that the action can be repeated with linear stretches in their overall duration (*i.e.*, change in global speed), in addition to of course their start (*i.e.*, temporal offsets). This is a reasonable assumption as we capture the same person repeating the same action. Second, we assume that the camera intrinsics are known, which can be easily satisfied by pre-calibrating the camera.

4. Calibration with The Human Body

Given the repeatable action videos, our first step is to estimate the camera viewpoints while temporally aligning the sequences. We achieve this spatio-temporal calibration by leveraging the human body, in particular, its joints as calibration targets. The human body is unique in that its joints and bones form a rigid structure that can be articulated. That is, the relative distances between the joints do not vary as they move in coordination. We exploit this fact to spatially calibrate the multiple views and also estimate the temporal stretch and offsets of the videos.

Spatio-Temporal Camera Calibration We first detect 2D human joints using OpenPose [10]. For each camera pair that captures, for instance, N frames of J 2D joints, we have $J \times N$ potential correspondences if the frames are temporally aligned. From these potential correspondences, we estimate the relative viewpoint of one camera to another by using the 5-point algorithm [35] on the 2D joints between the two sequences. We use random sampling consensus (RANSAC) [16] when computing the essential matrix, which implicitly takes care of the spatial variation of the repeated action across the two sequences.

We temporally calibrate the video sequences by estimating the offset and linear stretch of each video. As depicted in Fig. 3, we achieve this by explicitly evaluating the confidence defined as the number of inliers of the estimated essential matrix for each possible combination of temporal stretch a and offset b

$$\underset{a,b}{\operatorname{argmax}} \sum_{n}^{N} \sum_{j}^{J} C(\boldsymbol{x}_{n,k_1,j}, \boldsymbol{x}_{an+b,k_2,j}, c)\,, \tag{1}$$

where $\boldsymbol{x}_{n,k,j}$ is the 2D position of the j-th joint in the n-th frame n of the k-th video, and C denotes the number of 2D joint pairs whose Sampson distance [20] is lower than a threshold c for the essential matrix estimated with temporal stretch a and offset b. By evaluating the temporal alignment with sequence-to-sequence scores, we have far more correspondence points than frame-to-frame scores. In the experiments, we search for a from 0.7 to 1.3 and b from -1.5 seconds to 1.5 seconds.

The 2D search in temporal parameter space gives us the temporal calibration and initialization of the camera pose and the 3D joints. These frame-wise 3D joints have temporal noise, and the bone lengths are not necessarily temporally consistent. We refine 3D joints and camera poses with bundle adjustment [46]. To account for the spatial variation (*i.e.*, slight difference of each repetition), inconsistent bone length, and noisy 3D pose, we formulate bundle adjustment with additional regularization. Our objective function consists of a weighted sum of four terms: reprojection errors of the 3D joint E_{reproj}, temporal variances E_{bone} and left-right symmetry E_{sym} of the bone lengths, and temporal smoothness of the joint motions E_{smooth}

$$\begin{aligned} E_{\text{pose}} = & E_{\text{reproj}} + \lambda_{\text{bone}} E_{\text{bone}} + \\ & \lambda_{\text{sym}} E_{\text{sym}} + \lambda_{\text{smooth}} E_{\text{smooth}}\,. \end{aligned} \tag{2}$$

We define the reprojection error term E_{reproj} as the sum of the reprojection error of each of the J joints at all N frames in K viewpoints:

$$E_{\text{reproj}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} \left\| \Pi_k(\boldsymbol{X}_{j,n}) - \boldsymbol{x}_{n,k,j} \right\|^2\,, \tag{3}$$

where $\boldsymbol{X}_{j,n}$ denotes the 3D position of the jth joint at frame n, $\Pi_k(\cdot)$ projects a 3D point to the kth viewpoint, and $\boldsymbol{x}_{n,k,j}$ is the 2D position of the jth joint at frame n detected in the image of the kth viewpoint.

We define the variance of the bone length term E_{bone} as

$$E_{\text{bone}} = \sum_{b \in \mathcal{B}} \text{Var}_n(L_b), \quad L_b = \left\| \boldsymbol{X}_{j_b,n} - \boldsymbol{X}_{j'_b,n} \right\|^2\,, \tag{4}$$

where $b \in \mathcal{B}$ denotes each bone in the skeleton model, L_b denotes the length of bone b whose endpoints are j_b and j'_b, and $\text{Var}_n(\cdot)$ denotes the variance over time, *i.e.*, $n \in [1 : N]$. Similarly, we define the bone symmetry term E_{sym} as

$$E_{\text{sym}} = \sum_{n}^{N} \sum_{\langle b,b' \rangle \in \mathcal{S}} \left\| L_b - L_{b'} \right\|^2\,, \tag{5}$$

where $\langle b, b' \rangle \in \mathcal{S}$ denotes each of the symmetric bone pairs in the model such as, for example, the left and right forearms. The smoothness term evaluates the temporal continuity of the joint motion by the magnitude of its second derivative

$$E_{\text{smooth}} = \sum_{n=2}^{N-1} \sum_{j}^{J} \| -\boldsymbol{X}_{j,n-1} + 2\boldsymbol{X}_{j,n} - \boldsymbol{X}_{j,n+1} \|^2 . \quad (6)$$

Consistent 3D Skeleton Once the videos are spatially and temporally calibrated, we fit a 3D skeleton model (*i.e.*, representation of bone length and rotation) to the 3D joints in each frame to extract a consistent structural model of the human body across all videos and their frames. We employ an inverse kinetics model with a penalty term that prevents impossible joint angles. That is, we optimize the joints of the 3D skeleton $\hat{\boldsymbol{X}}_{j,n}(\boldsymbol{\theta}, \boldsymbol{\tau})$ parameterized by the joint angle vector $\boldsymbol{\theta}$ of the whole body and global translation parameter $\boldsymbol{\tau}$ by minimizing

$$E_{\text{IK},n}(\boldsymbol{\theta}, \boldsymbol{\tau}) = E_{\text{3D},n}(\boldsymbol{\theta}, \boldsymbol{\tau}) + \lambda_{\text{prior}} E_{\text{prior}}(\boldsymbol{\theta}) , \quad (7)$$

where $E_{\text{3D},n}(\boldsymbol{\theta}, \boldsymbol{\tau})$ denotes the sum of the distances between the skeleton joint $\hat{\boldsymbol{X}}_{j,n}(\boldsymbol{\theta}, \boldsymbol{\tau})$ and the corresponding joint $\boldsymbol{X}_{j,n}$ obtained from Eq. (2) for the n-th frame

$$E_{\text{3D},n}(\boldsymbol{\theta}, \boldsymbol{\tau}) = \sum_{j}^{J} \| \hat{\boldsymbol{X}}_{j,n}(\boldsymbol{\theta}, \boldsymbol{\tau}) - \boldsymbol{X}_{j,n} \|^2 . \quad (8)$$

$E_{\text{prior}}(\boldsymbol{\theta})$ is the reconstruction loss of a variational autoencoder pretrained on the AMASS dataset [32] and λ_{prior} weights its contribution. The reconstruction loss is evaluated with the poses inside a temporal window (5 frame) around the n-th frame. We initialize $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ by matching three joints, the neck and the left and right hip joints, which we experimentally found to always result in fast and stable convergence.

Finally, we reduce the remaining spatial discrepancies between reprojected 3D joints of the 3D skeleton model and the detected 2D joints in each viewpoint. Using the fit skeleton model, we absorb the reprojection errors by adjusting the skeleton pose for each viewpoint at each frame by evaluating a reprojection loss specific to the viewpoint k in addition to the inverse kinematics loss

$$\begin{aligned} E_{\text{IK},n,k}(\boldsymbol{\theta}, \boldsymbol{\tau}) = & E_{\text{3D},n}(\boldsymbol{\theta}, \boldsymbol{\tau}) + \lambda_{\text{reproj}} E_{\text{reproj},n,k}(\boldsymbol{\theta}, \boldsymbol{\tau}) \\ & + \lambda_{\text{prior}} E_{\text{prior}}(\boldsymbol{\theta}) , \end{aligned} \quad (9)$$

where, using camera projection Π,

$$E_{\text{reproj},n,k}(\boldsymbol{\theta}, \boldsymbol{\tau}) = \sum_{j}^{J} \| \Pi_k(\hat{\boldsymbol{X}}_{j,n}(\boldsymbol{\theta}, \boldsymbol{\tau})) - \boldsymbol{x}_{n,k,j} \|^2 . \quad (10)$$

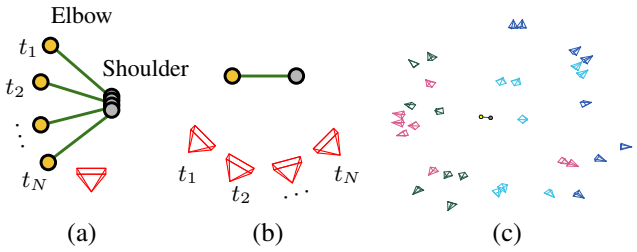

Figure 4. Bone-anchored camera transformation. A static camera capturing a bone in motion (a) can be transformed into a bone-anchored coordinate frame and collectively form a multiview system for each body part (b). These bone-anchored camera transforms virtually create a dense multview image capture for each body part (c) as shown in Fig. 2.

5. Clothed 3D Body Reconstruction

Given the spatially and temporally calibrated repeatable action videos and the consistent 3D skeleton model that explains the articulations in each frame of every video, we are now ready to recover the clothed body surface of the target person. We achieve this by virtually transforming the sparse viewpoints of each frame into a common coordinate frame with a rest-pose body by attaching and rotating each of the camera viewpoints to each of the bones. This effectively turns the original sparse multiview data into a dense calibrated multiview image set that can be used for conventional 3D geometric measurements. Note that due to the movements of the body, color consistency cannot be assumed and matching-based multiview methods are not applicable.

Bone-Anchored Camera Transformation Our key insight is to reinterpret an articulated 3D human body in front of a fixed camera view (Fig. 4 (a)) as virtual cameras anchored to static bones of a 3D human body in rest shape rotating around the body (Fig. 4 (b)). We anchor cameras to the 3D bones and independently transform the 3D bone into the rest pose along with the cameras that capture the bone. Virtually duplicated and transformed cameras are now capturing a single static 3D body in rest pose, just like in a multiview studio (Fig. 4 (c)).

This means that if we have K repeatable action videos each with N frames, each of the M body parts will have a maximum of $K \times N$ viewpoints observing its shape in the common coordinate frame of the rest pose. In other words, we will have a virtual multiview capture with $K \times N$ cameras for each of the M parts. The camera transformation can be easily computed as the inverse transformation of the 3D bone transformation matrix computed from 3D skeleton kinematics.

Body Part Visual Hulls Given the cameras transformed into a common coordinate frame, we can "measure" the

Figure 5. Sample images in S-RAD, a dataset consists of clothed 3D shapes of human performing repeatable actions. We use S-RAD for quantitative evaluations of our method.

body shape with well-established multiview methods. We adopt visual hull reconstruction, known as Shape from Silhouette, as it is robust to the appearance changes inherent to a body in action. Although the majority of the spatial and temporal variations have already been absorbed, visual hull reconstruction can result in over-carving by even one frame of incorrect virtual camera position that can arise from remaining pose errors or 2D joint detection errors. To combat this sensitivity, we automatically select the views used for visual hull reconstruction. We use the reprojection error between the 3D joints of the posed skeleton model as well as the confidence values returned by the 2D joint detector to select the views.

For each body part, we create a visual hull. Using selected images from transformed virtual cameras as depicted in Fig. 4 (c), we take an intersection volume of the body part silhouettes of the images. The visual hull is constructed based on a grid volume ($50\times50\times50$) defined around the body part of interest, and then converted into a mesh surface representation [28]. We first create a 3D point grid around the body part and project the points into each of the selected images of transformed cameras. 3D points that fall outside the silhouettes in any of the images are eliminated from the grid. By limiting the point grid to each body part region and also by only using selected images of each camera dramatically reduces the computation while ensuring detailed reconstruction of the body surface. Note that this multiview reconstruction has more or comparable number of cameras ($K\times N$) to existing multiview studios with actual cameras (*e.g.*, 3 to 100 cameras, according to [45]).

3D Body Surface Model We extract a dense body surface model from the visual hulls each representing a set of point samples of the body part surface in rest pose. For this, we fit a body surface model to the point samples. Unlike learned statistical models that can only represent body shapes within the range of training data, we can freely deform the surface model to fit the reconstructed visual hulls so that they capture the detailed clothed body shape. We directly optimize mesh vertex positions $v \in V$ of the SMPL model [31] by minimizing the Chamfer loss, the edge loss, the normal loss, and the Laplacian loss implemented in PyTorch3D [40]. The crucial point here is that we have actually measured the clothed body surface to which we fit the 3D body surface model.

6. Experimental Results

We introduce two new datasets to thoroughly evaluate the effectiveness of our method: the Real Repeatable Action Dataset (R-RAD) and the Synthetic Repeatable Action Dataset (S-RAD). These two datasets are complementary in that results on R-RAD demonstrate our method's performance in real-world settings, and S-RAD lets us quantitatively evaluate our method and compare with related methods on ground truth data.

Real Repeatable Action Dataset R-RAD consists of videos of people repeating natural repeatable actions captured with tripod-mounted 7 Sony DSC-RX0 cameras running at 120Hz in a studio (4 for reconstruction and 3 for testing). The original videos are temporally cut roughly into individual instances of repeated actions. To produce the randomness of this rough cut, we manually align the offset of the videos and shift the alignment by uniformly sampled random values from $[-0.5, 0.5]$ second. Randomness of the temporal stretch is introduced naturally from the repeated action. Silhouette is extracted with an existing background removal method [39]. The simultaneous multiview capture provides different variations of repeatable action videos and enables evaluation of reconstruction accuracy on views not used in the input set. In the following experiments, our method only uses one distinct view for each repeated instance.

Synthetic Repeatable Action Dataset Fig. 5 shows S-RAD, a dataset of repeatable action videos of synthetic humans. As it is nearly impossible to obtain accurate ground truth 3D surface of a person in action, such synthetic dataset becomes crucial for rigorous quantitative evaluation. We combine motion capture data with the MG-dataset [5] which contains textured mesh and its SMPL registration of clothed human. Our motion capture data were created from videos of people performing various repeatable actions captured from 7 viewpoints. We manually synchronized the videos and triangulate and optimize joints with a temporal smoothness term, and fit a 3D skeleton to each frame. This motion capture data provide 3 repeatable actions each with 8 repetitions. Each instance is spatially and temporally different as the action is performed by a human. We picked 3 models from the MG-dataset and rendered them with these repeated actions from a different viewpoint. The resulting dataset allows us to compare reconstructions with ground truth body surface geometry.

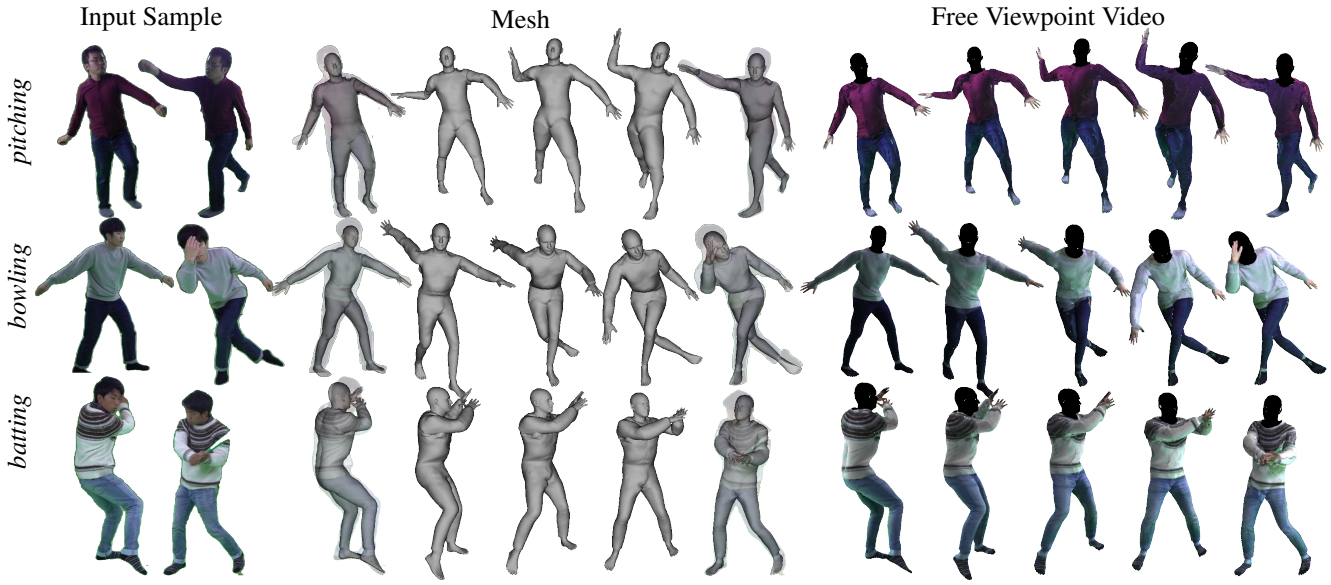

Figure 6. Qualitative reconstruction results of different subjects performing different actions on R-RAD. We show example frames from the input videos (left column), the reconstructed mesh (middle column), and the textured mesh (right column). In the middle column, left- and right-most images are rendered from the viewpoints in the input and overlaid on the input images. The other images are rendered from novel viewpoints. At the right column, we excluded head reconstruction, which is not articulate and out of the scope of our paper. Since our reconstruction is registered to the SMPL model, we can substitute these parts with the results estimated by another specialized method.

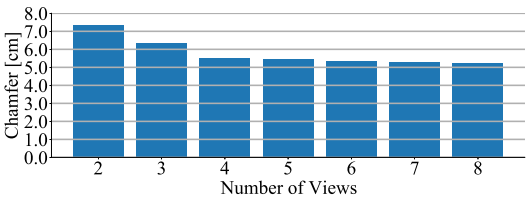

Figure 7. Ablation study of the number of views. The more views we have, the more accurate the reconstruction becomes. In practice, four viewpoints are sufficient for reasonable accuracy.

Table 1. Ablation study of pose and shape evaluated on all frames and all subjects in S-RAD. ↓ means smaller is better. The results show that the reconstruction improved with our method including temporal alignment in the spatio-temporal calibration and view adapted 3D pose optimization (Eq. 9).

Method	PA-MPJPE↓	Chamfer ↓
w/o temporal calibration	4.50	9.82
w/o view adaptation	2.54	6.14
Ours	**2.48**	**5.40**

Metrics We evaluate the reconstructed 3D pose with the Procrustes-Analysis Mean Per Joint Position Error in centimeters (PA-MPJPE). We evaluate the reconstructed shape with the Bidirectional Chamfer Distance (Chamfer) of the ground truth shape and result in centimeters.

Qualitative Evaluation We qualitatively evaluate the effectiveness of our method with R-RAD. Fig. 6 shows that our method successfully reconstructs body shapes in various actions and of different subjects. Our reconstructions accurately overlap the input images and are consistent with other views. Even though every action is a challenging pose sequence, the reconstruction result explains the input frames well.

Ablation Studies We evaluate the effect of varying the number of viewpoints for repeatable action video capture. As the number of viewpoints increases, the reconstruction error decreases. Fig. 7 shows the results. The results show that 4 viewpoints are sufficient, which is a very small number for multiview geometry reconstruction. Thanks to the bone-anchored camera transforms, the effective number of viewpoints are substantially increased by the number of frames of each instance sequence, enabling this reduction in physical viewpoints and thus practical and casual image capture.

We also conducted ablation studies to evaluate the importance of each step of our method. Table 1 shows the results. These results clearly show that the temporal and spatial variations in the repeatable actions are properly accounted for with our spatio-temporal calibration using the human body and all steps are important for recovering accurate body shape.

Figure 8. Qualitative comparison with other methods. We render the resulting shape from a novel view that does not exist in the input (test view). Our result is geometrically more consistent than other methods in that one shape matches the image from another viewpoint. Unlike single-image methods, our method does not suffer from inherent depth ambiguity.

Table 2. Quantitative comparison of reconstructed pose and shape (see text for metrics) on sparsely sampled 5 frames of all subjects in S-RAD. We compare our method with single image and single video methods. * indicates that the method uses their own clothed 3D shape as training data. Our method outperforms existing methods, while it does not require any training data of human body shape.

Method	Input	Chamfer ↓
PIFu* [42]	single image	7.48
DeepHuman* [52]		8.10
HMMR [25]	video	7.83
Ours	video (repeated action)	**4.77**

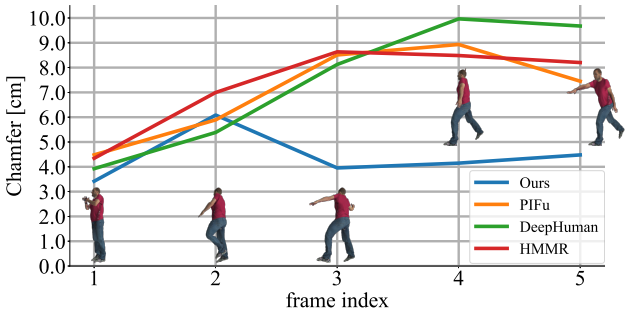

Figure 9. Evaluation of temporal consistency of the reconstructed clothed 3D human body shape. We compute the Bidirectional Chamfer Distance for sparsely sampled 5 frames of *pitching* action in S-RAD, to compare per-frame execution of single-image methods (PIFu [42], DeepHuman [52]), a video-based method (HMMR [25]), and our method (Ours). Single-image methods can reconstruct the shape when the person is in standing-like poses as they have less ambiguity and are likely closer to poses in their learning data (first frame). Our method reconstructs a consistent shape over time.

Comparison with Other Methods Fig. 8 shows qualitative comparison of our method with other methods. Our method exhibits geometrically accurate shape even from a different view that does not exist in the input videos. In contrast, other methods fail to infer or reconstruct the body shape that explains those in unseen views. Table 2 shows quantitative comparison of our method with other methods. Since there is no other method for shape reconstruction from repeatable action captured with a single camera, we compare our method with single-image methods. The results show that our method can accurately reconstruct the shape. Furthermore, our method does not rely on 3D ground truth clothed body shape data.

Evaluation on Temporal Consistency Fig. 9 shows the temporal change of the evaluation score of the estimated shape. Unlike single-image methods applied separately to every frame, our method can reconstruct a temporally consistent shape (i.e., a single surface model) that matches the ground truth articulated shape of every frame.

Outdoor Capture We apply our method to repeatable action videos of a person pitching a ball outdoors. As shown in Fig. 1, our method works with a single camera in an outdoor setting.

7. Conclusion

In this paper, we tackled the challenging problem of consistent 3D clothed human body shape recovery from casually taken images by fully leveraging the repeatability of actions and the human articulated body for spatio-temporal calibration and multivew view expansion. The experimental results show that the proposed method successfully reconstructs a consistent clothed body shape that matches all frames in the video. Each of the key steps builds on well-established multiview geometry concepts. The main contribution of our framework lies in the very idea of turning a casually taken video sequence into a fully calibrated multiview data to achieve body part-based reconstruction so that these established geometric methods can be exploited to arrive at an actual measurement of the unique clothed body shape of the person in the video. Unlike learning-based methods, our method produces a clothed, consistent 3D human model based on geometric measurements. We believe our method complements learning-based methods and clearly demonstrates what purely geometric approaches can still offer.

Acknowledgement This work was in part supported by JSPS KAKENHI 17K20143, JST PRESTO JPMJPR1858, and RIKEN Guardian Robot Project.

References

[1] Michael J. Black Ahmed A. A. Osman, Timo Bolkart. Star: Sparse trained articulated human body regressor. In *Proc. ECCV*, Aug. 2020. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *Proc. CVPR*, pages 8387–8397, 2018. 3

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. In *Proc. ACM SIGGRAPH*, pages 408–416. 2005. 2

[4] Serge Belongie and Josh Wills. Structure from periodic motion. In *Spatial Coherence for Visual Motion Analysis*, pages 16–24, 2006. 2, 3

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3d people from images. In *Proc. ICCV*. IEEE, Oct 2019. 6

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. ECCV*, pages 561–578. Springer, 2016. 3

[7] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proc. CVPR*, 2014. 2

[8] Eugene Borovikov and Larry Davis. A distributed system for real-time volume reconstruction. In *Proc. CAMP*, pages 183–189, 2000. 2

[9] Andrea Bottino and Aldo Laurentini. A silhouette-based technique for the reconstruction of human movement. *CVIU*, 83:79–95, 2001. 2

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 4

[11] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Transactions on Graphics (TOG)*, 22(3), 2003. 2

[12] Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. Silhouette-based object phenotype recognition using 3D shape priors. In *Proc. ICCV*, pages 25–32. IEEE, 2011. 3

[13] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *Proc. ACM SIGGRAPH*, pages 1–10. 2008. 2

[14] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: Neural articulated shape approximation. In *Proc. ECCV*, Aug. 2020. 2

[15] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *Proc. ECCV*, 2020. 3

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[17] Nicola Garau, Francesco G. B. De Natale, and Nicola Conci. Fast automatic camera network calibration through human mesh recovery. *Journal of Real-Time Image Processing*, (17):1757–1768, 2020. 3

[18] Peng Guan, A. Weiss, A. O. Bãlan, and M. J. Black. Estimating human shape and pose from a single image. In *Proc. ICCV*, pages 1381–1388, 2009. 3

[19] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *Proc. CVPR*, pages 5052–5063, 2020. 3

[20] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 4

[21] Jean-Marc Hasenfratz, Marc Lapierre, Jean-Dominique Gascuel, and Edmond Boyer. Real-Time Capture, Reconstruction and Insertion into Virtual World of Human Actors. In *Vision, Video and Graphics*, pages 49–56. Eurographics, 2003. 2

[22] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *Proc. CVPR*, pages 3093–3102, 2020. 3

[23] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, pages 34–47, 1997. 2

[24] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, pages 7122–7131, 2018. 2, 3

[25] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proc. CVPR*, pages 5614–5623, 2019. 8

[26] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. CVPR*, pages 4501–4510, 2019. 3

[27] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the People: Closing the loop between 3D and 2D human representations. In *Proc. CVPR*, 2017. 3

[28] Victor Lempitsky. Surface extraction from binary volumes with higher-order smoothness. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1197–1204. IEEE, 2010. 6

[29] V. Leroy, J. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proc. ICCV*, pages 3113–3122, 2017. 2

[30] Xiu Li, Hongdong Li, Hanbyul Joo, Yebin Liu, and Yaser Sheikh. Structure from Recurrent Motion: From rigidity to recurrency. In *Proc. CVPR*, pages 3032–3040, 2018. 2, 3

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *Proc. ACM SIGGRAPH Asia*, pages 248:1–248:16, 2015. 2, 6

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proc. ICCV*, Oct 2019. 5

[33] Saied Moezzi, Li-Cheng Tai, and Philippe Gerard. Virtual view generation for 3D digital video. *IEEE Multimedia*, pages 18–26, 1997. 2

[34] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *Proc. ICCV*, pages 900–908, 2015. 2

[35] David Nistér. An efficient solution to the five-point relative pose problem. *TPAMI*, 26(6):756–770, 2004. 4

[36] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural Body Fitting: Unifying deep learning and model based human pose and shape estimation. In *Proc. 3DV*, pages 484–494. IEEE, 2018. 3

[37] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3D human modeling. *Pattern Recognition*, 2017. 2

[38] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3):207–232, 2004. 3

[39] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. volume 106, page 107404, 2020. 6

[40] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[41] Evan Ribnick and Nikolaos Papanikolopoulos. 3D reconstruction of periodic motion from a single view. *IJCV*, 90(1):28–44, 2010. 2, 3

[42] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, 2019. 3, 8

[43] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proc. CVPR*, June 2020. 1, 3

[44] Thomas Schöps, Torsten Sattler, Christian Häne, and Marc Pollefeys. Large-scale outdoor 3D reconstruction on a mobile device. *CVIU*, 157:151–166, 2017. 3

[45] Jonathan Starck, Atsuto Maki, Shohei Nobuhara, Adrian Hilton, and Takashi Matsuyama. The multiple-camera 3-D production studio. *TCSVT*, 19(6):856 –869, 2009. 2, 6

[46] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 4

[47] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proc. ECCV*, pages 20–36, 2018. 3

[48] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 2

[49] Grorge Vogiatzis, Carlos Hernandez, Philip .H.S. Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *TPAMI*, 29(12):2241 –2246, 2007. 2

[50] Kwan-Yee Kenneth Wong and Roberto Cipolla. Reconstruction of sculpture from its profiles with unknown camera positions. *IEEE Transactions on Image Processing*, 13(3):381–389, 2004. 3

[51] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proc. CVPR*, 2019. 3

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *Proc. ICCV*, October 2019. 8