Outdoor Markerless Motion Capture with Sparse Handheld Video Cameras

Yangang Wang, Yebin Liu, Xin Tong, Qionghai Dai, and Ping Tan, *Member, IEEE*

Abstract—We present a method for outdoor markerless motion capture with sparse handheld video cameras. In the simplest setting, it only involves two mobile phone cameras following the character. This setup can maximize the flexibilities of data capture and broaden the applications of motion capture. To solve the character pose under such challenge settings, we exploit the generative motion capture methods and propose a novel model-view consistency that considers both foreground and background in the tracking stage. The background is modeled as a deformable 2D grid, which allows us to compute the background-view consistency for sparse moving cameras. The 3D character pose is tracked with a global-local optimization through minimizing our consistency cost. A novel L_1 motion regularizer is also proposed in the optimization to constrain the solution pose space. The whole process of the proposed method is simple as frame by frame video segmentation is not required. Our method outperforms several alternative methods on various examples demonstrated in the paper.

Index Terms—Markerless motion capture, handheld video cameras, model-view consistency

1 Introduction

Markerless motion capture records the 3D motion of an actor without marker or sensor suits. It has a wide range of applications including human computer interaction, surveillance, gait analysis, and visual special effects in games and movies. Despite recent advances that enable people to capture the high quality motions via specialized device setups [1], [2], [3], [4] or depth cameras [5], [6], [7], [8], [9], [10] in an indoor environment, outdoor markerless motion capture is still a challenge task where the character moves freely in a open space with uncontrolled background and illumination. A few pioneering works [11], [12], [13], [14], [15] explored outdoor markerless motion capture. However, most of these methods still need substantial works on establishing the capture settings.

To maximize the data capture flexibility, we advocate to study outdoor markerless motion capture with sparse handheld video cameras. We allow the handheld cameras to follow the moving character and be carried by one or several people. In the simplest setting, it only involves two mobile phone cameras following the character. This simple setup has the potential to enormously broaden the diversity and richness of motions that can be captured, such as skateboarding, basketball, or even ski. At the same time, it also poses a big challenge for pose estimation algorithm due to the large pose ambiguities caused by sparse views, uncontrolled (and possibly variant) illumination, and inaccurate camera calibration.

In this paper, we exploit the generative approach [3], [16] to estimate the pose via an analysis-by-synthesis approach. Specifically, we model the character mesh as a surface triangle mesh driven by a skeleton [3] and search for a skeleton pose that maximizes the model-view consistency for all views.

Typical generative methods measure the model-view consistency with the color differences between the projected character model and input images, which is plagued by two difficulties, especially for sparse handheld cameras. First, different body parts might have similar color to confuse the model to image matching. This is exemplified in Figs. 1a and 1b. The arms of the character in (a) have similar color as the torso. As a result, a wrong pose where the two arms are in front of the torso produces strong model-view consistency. Even if the arms have different colors as the torso, the arm itself can still cause similar problems as shown in (b). In this case, the arm incorrectly bends forward, but the model-view consistency is high as the forearm has uniform color. Second, the background might have similar color to some body parts. Wrong matches between model and image might produce good consistency. This is exemplified in Fig. 1c, where the brick pavement has similar color as the arms and accidently produces better consistency.

Some existing methods [17], [18] rely on a relatively large set of cameras (typically more than 4) to solve the pose ambiguities. For the example in Fig. 1a, if a camera sees the character from the side, it will produce strong penalty for the arms in front of torso. Other methods [1], [3] rely on a background model to segment the foreground out before the pose estimation to address the challenges. By requiring the projected mesh to fill-in the segmentation mask, the

- Y. Wang and X. Tong are with the Microsoft Research Asia, Microsoft Research, Beijing 100083, China.
E-mail: ygwangthu@gmail.com, xtong@microsoft.com.
- Y. Liu and Q. Dai are with the Department of Automation, Tsinghua University, Beijing 100083, China.
E-mail: liuyebin@mail.tsinghua.edu.cn, qhdai@tsinghua.edu.cn.
- P. Tan is with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 16S, Canada. E-mail: pingtan@sfu.ca.

*Manuscript received 31 May 2016; revised 28 Mar. 2017; accepted 3 Apr. 2017. Date of publication 12 Apr. 2017; date of current version 28 Mar. 2018.
Recommended for acceptance by Y. Yu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2017.2693151*

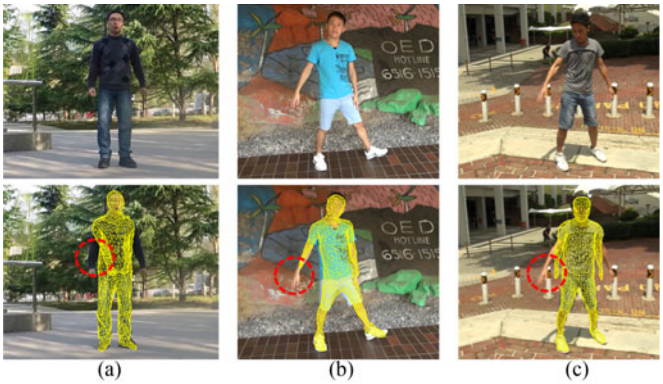


Fig. 1. Some wrong poses associated with high model-view consistency score: (a) arms with similar color as torso; (b) arms with uniform color; (c) background with similar color as foreground (arms).

challenging cases in Figs. 1a, 1b, and 1c can be safely handled. However, it is still an open problem to automatically obtain fine segmentation masks. Even with iterative segmentation (with background model) and pose tracking (foreground model) [12], [19], [20], video segmentation is tedious, error-prone [16] and sometimes manual intensive.

To tackle these pose ambiguities, we present a novel model-view consistency cost that considers both the foreground and background under the setting of sparse handheld cameras. Our observation is that a correct pose can produce high consistency for both foreground-view and background-view at the same time. Our foreground model is built at each mesh triangle. Different from previous background modeling for markerless motion capture [9], [18], [21], [22], we model the background as a deformable 2D grid, which allows us to compute a background-view consistency for sparse moving cameras. With such new model-view consistency, our optimization solves for the human poses by parsing the image pixels according to a normalized foreground and background consistency cost. For each pixel in the image, we compute its consistency to both foreground and background simultaneously. As a result, the wrong poses as in Fig. 1 can be penalized as some pixels in the character are mistaken as in background. To deal with illumination variations over time, we also introduce a view-view consistency cost bridged by our character model. A novel L_1 motion regularizer is utilized in the optimization to enforce temporal pose coherence.

By combining these novel designs, our method efficiently solves the pose ambiguities and robustly reconstructs the 3D poses from sparse videos captured by handheld cameras. The whole process of our method is simple as frame by frame video segmentation is not required. We evaluate our method with variant video sequences and illustrate the advantages of our method to other alternative solutions. We also demonstrate the proposed method with as few as two *uncalibrated* and *unsegmented* video sequences, which brings the markerless motion capture to a stage much closer to applications.

2 Related Work

Many different motion capture algorithms have been proposed. A complete survey is beyond the scope of this work. Here we mainly discuss some recent generative motion capture methods, since our method belongs to this category.

Generative Methods. Deform a character model with simple shape primitives [23], [24] to estimate the skeleton pose by maximizing the model-view consistency. Some recent methods even model the detailed 3D deforming surface [2], [25], [26], [27], [28] of the actor. The energy function can be solved by local [29], [30] or global optimization [24], [31], or a combination of the two [3]. Recently, [32] presented a new scene representation that enables an analytically differentiable closed-form formulation of surface visibility, which yields smooth, analytically differentiable, and efficient to optimize pose similarity energies with rigorous occlusion handling, fewer local minima, and experimentally verified improved convergence of numerical optimization.

There is a clear trend to reduce capture setup complexity. Early methods typically require controlled studios with a large number of cameras [33], [34]. Tresadern and Reid [35] pioneered this direction using two uncalibrated and unsynchronized cameras, and captured rough skeleton positions. Elhayek et al. [17] employed a dozen of unsynchronized cameras and achieved strong results in indoor environments. Hasler et al. [12] used handheld cameras in outdoor scenes by exploiting silhouette cues. Silhouette involves tedious video segmentation and also requires cameras to have a large view span, which is difficult when capturing in a cluttered scene. Elhayek et al. [36] solved camera calibration and pose estimation simultaneously. Their method is demonstrated with at least five cameras, among them at least two are fixed, to resolve the ambiguity between the camera ego-motion and character motion. Wei et al. [37] studied motion tracking from a monocular video with manual interventions. Under the assumption of fixed distant illumination, Wu et al. [4] exploited 3D points reconstructed from a stereo rig for performance capture. However, their local optimization and fixed distant illumination model is only demonstrated in indoor scenes. Ganapathi et al. [9] utilized the depth cameras and extended the iterative closest points (ICP) objective by modeling the constraint that the observed subject cannot enter free space to obtain real-time human pose tracking results.

We aim at outdoor markerless motion capture with sparse handheld cameras. Our formulation considers both the foreground and background consistency, though background modeling has been used in some works. [18] constructs a SOG representation for each input image (at both foreground and background regions). However, they only compute foreground consistency while ignoring the background consistency in the tracking stage, thus their method need more than 5 views. Recently, Loper and Black [22] propose a general differentiable renderer named OpenDR for analysis-by-synthesis approach and build models for both the foreground and background. Nevertheless, their example on pose tracking yet only computes the foreground consistency and their method is only demonstrated on a RGB-D input for static background.

Our work is related to PoseCut [19], which solves the optimal skeleton by minimizing the graph-cut energy. In comparison, we discard the graph-cut based segmentation. Graph-cut is time-consuming and cannot guarantee the segmented shape to be a valid character silhouette, though soft shape priors are used in [4], [19]. Our foreground is directly determined by the deformed mesh model, such that we use sampling-based method to directly optimize the skeleton pose.

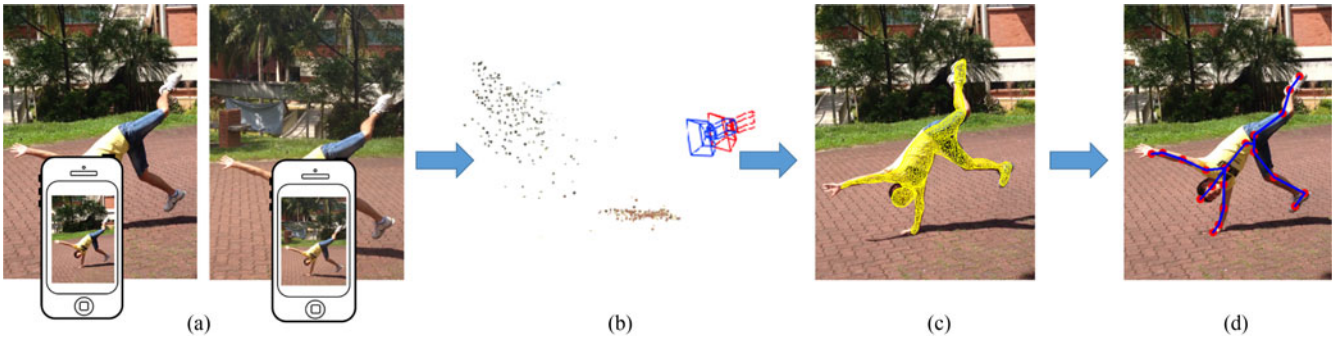


Fig. 2. System pipeline. (a) Our system takes multiple synchronized videos captured by handheld cameras as input. (b) We adopt the CoSLAM [Zou and Tan 2013] method to calibrate all cameras in 3D space. (c) Our motion tracking system estimates the skeleton motion with a sample-based method. (d) We use a gradient-based method to refine the skeleton pose.

Discriminative Methods. [5], [6] have demonstrated real-time performance and enabled enormous applications with a depth camera. [10] proposed a novel training objective which leads to considerably more accurate correspondences with far fewer training images. Skeleton motion of multiple human bodies can also be track using the discriminative method under static multi-camera setup [38] or even hundreds of RGB cameras [39].

Hybrid Methods. Combine the strength of discriminative methods and generative methods. Some recent works [8], [40], [41] combine a discriminative method with the local optimization from generative methods to improve the results. Most recently, Elhayek et al. [15] combined the convolution neural network based skeleton estimation with the Sum-of-Gaussian [18] based pose tracking, and achieved surprisingly good results with only 2-3 cameras. Bogo et al. [42] also used the convolution neural network to predict the 2D body joint locations and then fitted the SMPL [43] (a statistical body shape model) to the 2D joints. They minimized an objective function that penalizes the error between the projected 3D model joints and detected 2D joints. Our novel formulation of model-view consistency might be adopted in these hybrid methods to further improve their performance.

Wearable Sensors. Such as IMUs and accelerometers [14], [44], [45] can also be used to facilitate outdoor motion capture. However, these techniques still rely on a motion database to infer human pose from the noisy acceleration data. They are more suitable for repeatable and distinctive motions. Commercial inertial motion capture systems (e.g., Xsens MVN, www.xsens.com) can capture body motion in outdoor scenes with a well-equipped compact suits. In comparison, our techniques provides an extremely flexible capture setup without wearable markers or sensors.

3 Preliminaries

The pipeline of our system is summarized in Fig. 2. Our inputs are multiple video streams captured by a few (2-3) handheld cameras. In all our examples, the resolution of input videos is 960×540. The video streams are synchronized by flashing a light such that the synchronized frames can be detected to calculate the integer frame shifts between cameras, though voice synchronization in [12] might also be used. In the following, we assume all cameras are synchronized from the first frame to simplify notation.

We take the CoSLAM [46] algorithm to estimate the camera parameters. The sequential tracking framework is adopted and we solve the markerless motion capture by energy optimization. For each frame, we maximize the normalized character-view and background-view consistency to solve for the skeleton pose. Benefits from the proposed appearance model, our system can robustly recover the human pose from sparse handheld cameras without the frame by frame video segmentation. We use a sampling-based global optimization method to explore the large skeleton pose space, and refine the result by a local gradient descendent search. The global optimization can avoid temporal error accumulation during pose tracking.

3.1 Character Model

We adopt the analysis-by-synthesis approach [47] to estimate 3D skeleton poses. Our character model includes a triangle mesh $\mathcal{M}$ registered with a kinematic skeleton χ. The mesh model $\mathcal{M}$ is prepared interactively according to multi-view images of the target character in a fixed standing pose, with public tools such as VisualSFM [48] and multi-view stereo systems. $\mathcal{M}$ has about 5,000 triangles. It is noted that the recent publicly available SMPL [43] (a statistical body shape model) can be used to obtain a better rigged mesh model $\mathcal{M}$ from the multi-view reconstructed point cloud.

Fig. 3 shows our skeleton model χ with 39 DoFs (6 for the global motion and 33 for the joints). The skeleton χ is parametrized by $(\theta_0\hat{\xi}_0, \Theta)$, where $\theta_0\hat{\xi}_0$ is the twist coordinate encoding the global translation and rotation of the whole human body, and $\Theta = \{\theta_1, \ldots, \theta_n\}$ is the vector of joint angles.

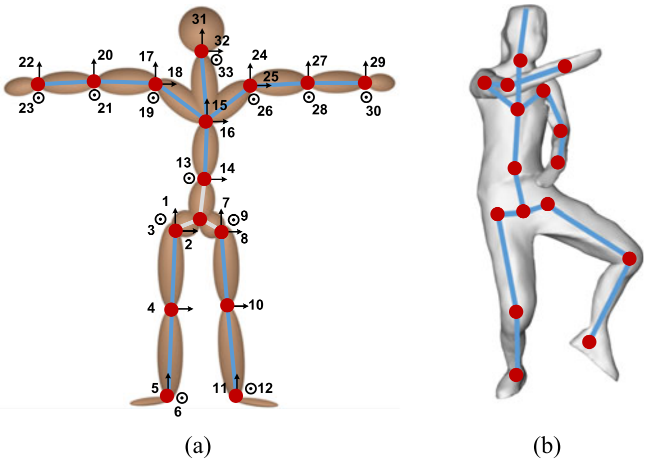


Fig. 3. (a) our skeleton model with joints and DoFs; (b) a character mesh $\mathcal{M}$ deformed according to the skeleton pose.

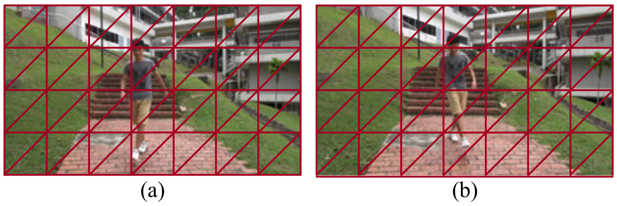


Fig. 4. We build the background model with a deformable uniform mesh in the selected frames. (a) and (b) are the selected two frames with uniform 2D grid mesh. A mixture Gaussian model is built within each triangle of this mesh excluding the character regions.

This skeleton χ is automatically registered with $\mathcal{M}$ and surface skinning weights are automatically computed as described in [49]. Thus, the coordinate of a mesh vertex $\mathbf{v}$ under the skeleton pose χ can be computed by linear blend skinning as,

$$\mathbf{v}(\chi) = \left(\sum_{i \in \mathcal{B}} \omega_i T_i\right) \mathbf{v}_0. \tag{1}$$

Here, $\mathbf{v}_0$ is the model vertices on the initial pose, $\mathcal{B}$ is the set of joints, ω_i is the binding weight between $\mathbf{v}$ and the ith joint and $\sum_{i \in \mathcal{B}} \omega_i = 1$. The transformation T_i defines the motion of the ith joint, which is

$$T_i = e^{\theta_0 \hat{\xi}_0} \cdot \prod_{j \in \mathcal{P}_i} e^{\theta_j \hat{\xi}_j}, \tag{2}$$

where $\mathcal{P}_i$ includes all the parent DoFs of the ith joint, and $\hat{\xi}_j$ is a corresponding twist matrix. The Jacobian matrix of Equation (2) has a special structure, which can be easily obtained from its adjoint transformation. More details can be found in the chapter 2 and 3 of [50].

3.2 Appearance Model

Background. We propose a novel background model with a deformable 2D grid of triangles as shown in Fig. 4, where each triangle contains a mixture of K Gaussians to represent its pixel distribution in the RGB color space (K is 5 in this paper). The background model does not consider the color information from multiple frames, otherwise we build the background model on each of the selected video frames, $\Gamma = \{\gamma_1, \ldots, \gamma_n, \ldots, \gamma_N\}$, where N is the number of selected frames (N is from 1 to 5 in all experiments of this paper). Generally, the video frames are uniformly selected from the whole video sequences.

Considering the color variations among different cameras, the background model is also built for each camera. For the triangle $\triangle$ in frame γ_n of the mth camera, we adopt K-means to obtain K clusters and build a Gaussian model for each cluster in the RGB space. The model distance of a pixel with color $\mathbf{c}$ is computed as,

$$\psi_m^{\triangle,\gamma_n}(\mathbf{c}) = \min_{k \in K}\left((\mathbf{c} - \mu_k)^\mathsf{T} {\sum}_k^{-1} (\mathbf{c} - \mu_k), \tau\right). \tag{3}$$

Here, μ_k and $\sum_k$ are the mean and variance of the Gaussian model, respectively. τ is a large constant barrier for color values, which is set to 1 for color values between $[0, 1]$ and fixed in all experiments in this paper.

With the proposed background model, we can obtain the pure background without human characters by warping the selected frames and performing the image inpainting. Fig. 5 shows some foreground character removal results.

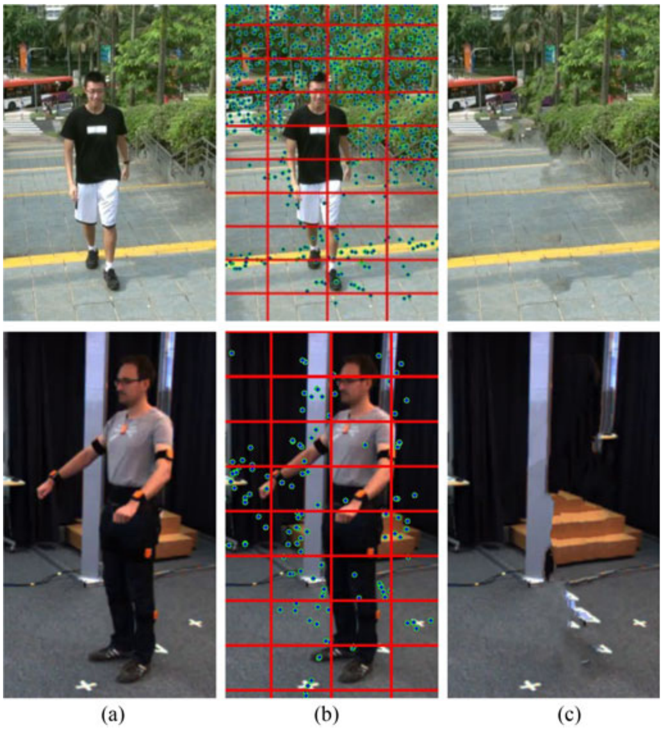


Fig. 5. Background warping and foreground removal results. (a) are the original images, (b) are the background warping, and (c) are the foreground removal results.

Foreground. The foreground color model is directly built at each triangle of the mesh $\mathcal{M}$. For this purpose, we adopt the method in [3] to register the mesh $\mathcal{M}$ at the selected frames Γ. At each mesh triangle $\triangle$, we collect all its covered pixels from the frames in Γ, and utilize the same strategy as background to build the color model. The model distance is similar as Equation (3) except that the foreground model distance does not depend on the selected frame index γ_n. It should be noted that not all mesh triangles are visible in the selected video frames. Thus the proposed foreground color model is not watertight for all the mesh triangles.

For the selected frames Γ, we utilize [3] to register the mesh $\mathcal{M}$ with the observed images and obtain the foreground color model. It is noted that image silhouette is needed for human pose estimation of [3], the foreground masks of Γ frames are manually segmented beforehand. Besides, we found the method in [3] is not robust under the scenario of 2-3 cameras. We manually correct about 6 DoFs (39 DoFs in total) especially when there exist self-occlusions. A possible solution is to integrate discriminative methods with [3] and improve the complexities of initializing our appearance model. Furthermore, considering the illumination variations of the video sequences, manually select the foreground and background frames would make our appearance models more accurate.

4 Pose Estimation

Our pose estimation starts from a rough registered skeleton at the first frame and tracks the skeleton in a frame-by-frame manner. At each frame, we first adopt a sampling-based global optimization to find a rough estimation. We then take a gradient descendent method to refine the pose locally. In the whole process, we deliberately remove the silhouette constraints to avoid tedious video segmentation.

4.1 Sample-Based Pose Optimization

For each sampled pose χ, we deform the template mesh and project it into images via camera parameters. An energy function is then computed from all pixels in the image plane. In order to accelerate the computation, we bound the computed pixels in a rectangle, which is estimated from the projected silhouette of previous frame and the precomputed average optical flow of the character pixels.

Note the mesh model $\mathcal{M}(\chi)$ should be first computed to evaluate χ. This involves updating each vertex according to Equation (1). (We often omit the dependency on χ when referring to a vertex $\mathbf{v}$ or $\mathcal{M}$ for notation simplicity.) For better computation efficiency, we take the Taylor expansion of $\mathbf{v}(\chi)$ and compute the vertex coordinate as,

$$\mathbf{v}(\chi + \nabla\chi) = \mathbf{v}(\chi) + J\nabla\chi. \tag{4}$$

Here, J is a Jacobi matrix [50]. In this way, we can compute $\mathcal{M}$ efficiently when continuously updating χ.

The Energy Function of sample-based pose optimization contains a model-view consistency and a view-view consistency as the following,

$$E_{opt}(\chi) = \sum_m E_m^{mv}(\chi) + \lambda \sum_{m,n} E_{m,n}^{vv}(\chi). \tag{5}$$

The combination weight λ is fixed as 0.3 in all of our experiments. Here, m, n are the indices for cameras. The model-view consistency $E_m^{mv}(\chi)$ interprets each pixels as foreground or background according to the proposed appearance model. The view-view consistency $E_{m,n}^{vv}(\chi)$ is an implicit 3D constraint, which can alleviate the appearance variation caused by temporal illumination changes.

The Model-View Consistency evaluates a cost at each pixel and sums it over all pixels. Specifically, we evaluate the distance of each pixel to the foreground character and the background as described in Equation (3). We denote the model distances of pixel $\mathbf{p}$ to the background and foreground as ψ_m^b and ψ_m^f, respectively. If the pixel $\mathbf{p}$ is covered by the projected character mesh $\mathcal{M}$ under the sampled pose χ, its consistency cost is computed as $\psi_m^f/(\psi_m^b + \psi_m^f)$. Otherwise, its consistency is evaluated as $\psi_m^b/(\psi_m^b + \psi_m^f)$. Briefly, this term has the following formulation,

$$E_m^{mv}(\chi) = \frac{1}{M_1} \sum_{\mathbf{p}} \begin{cases} \psi_m^b/(\psi_m^b + \psi_m^f), & \mathbf{p} \notin \mathbf{S}; \\ \psi_m^f/(\psi_m^b + \psi_m^f), & \mathbf{p} \in \mathbf{S}. \end{cases} \tag{6}$$

Here, M_1 is the number of pixels in the bounded rectangle and $\mathbf{S}$ is the set of pixels covered by the projected mesh $\mathcal{M}$. The model-view consistency cost computation is illustrated in Fig. 6.

The computation of ψ_m^b is frame-dependent, where we use the closest frame to the current frame in Γ, denoted as $\widehat{t}$, to compute ψ_m^b. The meaning of Γ is described in Section 3.2. Mathematically, $\widehat{t} = \arg\min|\gamma_n - t|, \gamma_n \in \Gamma$, where t is the index of current frame. We warp the $\widehat{t}$th frame to the current frame by the as-similar-as-possible method in [51]. After warping, if a pixel $\mathbf{p}$ is covered by a triangle $\triangle$ on the background in the $\widehat{t}$th frame, we use the triangle color model $\psi_m^{\triangle,\widehat{t}}$ to compute the background model distance of the pixel $\mathbf{p}$. Otherwise, we find the nearest background triangle in the warped image to compute the background model distance. Since the background model distance ψ_m^b does not vary with sampled poses, we can pre-compute ψ_m^b to accelerate the computation during the sampling based optimization.

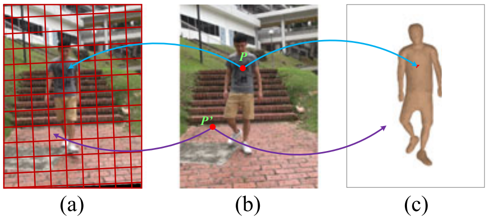


Fig. 6. The model-view consistency is evaluated on all image pixels. We evaluate the distance of each pixel to the foreground character and the background as described in Equation (3) and sums the cost over all pixels to obtain the model-view consistency.

As for the foreground model distance ψ_m^f, we compute it according to the projected mesh $\mathcal{M}$. Specifically, for a pixel $\mathbf{p}$ covered by a triangle $\triangle$ from $\mathcal{M}$, we compute ψ_m^f with the color model of $\triangle$. If $\mathbf{p}$ is not covered by any triangles, its ψ_m^f is set to τ as described in Section 3.2. It should be noted that not all mesh triangles have color models. We also set τ as the foreground model distances for those pixels covered by such triangles.

The View-View Consistency is computed only for pixels covered by the projected mesh $\mathcal{M}$. Suppose pixel $\mathbf{p}_m$ in camera m is covered by a triangle $\triangle_1$ on the projected mesh $\mathcal{M}$. We obtain its 3D position by intersecting $\triangle_1$ with the view ray passing through $\mathbf{p}_m$. We then compute the projected 2D position of this 3D point in camera n via its parameters, which is denoted as $\mathbf{p}_n$. It is noted that $\mathbf{p}_m$ and $\mathbf{p}_n$ may be covered by different triangles due to mesh self-occlusions. This mesh triangle coverage can be easily checked though rendering procedure. Suppose $\mathbf{p}_n$ is covered by $\triangle_2$, we compute the geodesic distance $Geo(\triangle_1, \triangle_2)$ between $\triangle_1$ and $\triangle_2$, which is the shortest path from $\triangle_1$ to $\triangle_2$ on the mesh $\mathcal{M}$. One possible condition is that $\mathbf{p}_n$ may not be covered by any triangles of the projected character mesh, where the geodesic distance is computed as infinity. If the geodesic distance is less than a threshold (in all our experiments, this threshold is set as five triangles), $\mathbf{p}_m$ and $\mathbf{p}_n$ are considered in correspondence and we compute the view-view consistency cost by their euclidean distance in RGB space. Otherwise, if the geodesic distance is larger than the threshold, the view-view consistency cost is set to τ. (please refer to Equation (3) for τ)

We use $c_n^t(\mathbf{p}_m)$ to denote the euclidean distances between $\mathbf{p}_m$ and $\mathbf{p}_n$. The view-view consistency term is finally defined as,

$$E_{m,n}^{vv}(\chi) = \frac{1}{M_2} \sum_{\mathbf{p} \in \mathbf{S}} \begin{cases} \min(c_n^t(\mathbf{p}_m), \tau), & Geo(\triangle_1, \triangle_2) \leq 5; \\ \tau, & otherwise. \end{cases} \tag{7}$$

Here, M_2 is the number of pixels covered by $\mathcal{M}$. The view-view consistency cost computation is illustrated in Fig. 7.

The Optimization of Equation (5) is solved by a sample-based method. Our skeleton pose χ has 39 degree-of-freedoms (DOFs). Directly sampling in this 39D space requires a large set of samples to guarantee good results. For example, suppose we sample each dimension independently,

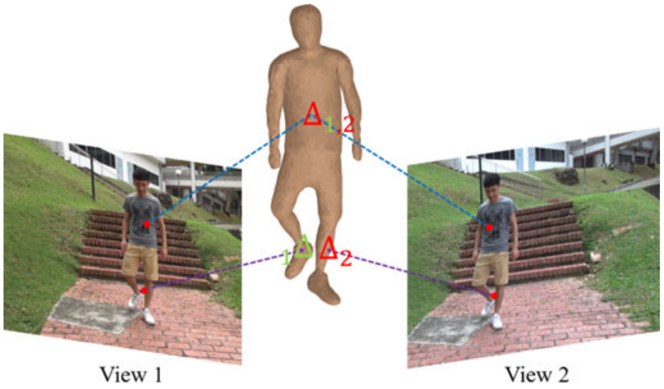


Fig. 7. The view-view consistency is computed only for pixels covered by the projected mesh. We use the projected mesh triangle index to build the pixel correspondences across different views and compute the consistency cost by their euclidean distances in RGB space.

and take equal probability to increase or decrease the value in each dimension. In this way, in each dimension, there are 50 percent chance to create a sample that is closer to the optimal result. This naïve sampling scheme will require generating and evaluating 2^{39} samples at each iteration to guarantee good performance. For efficiency consideration, we take a divide and conquer approach to solve the skeleton poses of the four limbs independently.

Specifically, we divide the character to torso and four limbs. We first solve the global rigid motion and parameters of joints on the torso. After that, we solve the joint parameters of each limb respectively. In this hierarchical approach, each time the DOFs is kept at $6 \sim 7$ and a much smaller set of samples are sufficient. In the sampling-based optimization, Equation (5) is minimized via the *interacting simulated annealing* (ISA) techniques [31]. We randomly sample $2^n + P$ (n is the DOFs, $P = 40$ in our experiments) skeleton poses based on the result at previous iteration. Each sample is weighted by the energy E_{opt} of Equation (5). New skeleton poses are generated through the re-sampling according their weights and mutation procedures. We repeat this iterative re-sampling and mutation 25 times, and compute the final result as a weighted average of the last set of skeleton poses.

*Initialization.*We initialize the sampling-based optimization by the skeleton pose from the previous frame. To facilitate quick convergence, we adopt the optical flow based tracking method in [4] to improve the initialization, which gives faster convergence rate (e.g., smaller iteration times).

4.2 Gradient Descendent Pose Refinement

The sample based pose optimization enjoys small drifting error during pose tracking. But the randomly generated samples are often suboptimal, e.g., the captured skeleton motion suffers from significant temporal jitter. Therefore, we apply a local gradient descendent based method to refine the skeleton pose.

For better efficiency, we adopt a simplified energy term than Equation (5) as the following,

$$E_{ref}(\chi) = E^{data}(\chi) + \lambda_{reg} E^{reg}(\chi). \tag{8}$$

The term E^{data} is a data term measuring how well the deformed pose fits to the observation. The term E^{reg} is a smoothness prior. The coefficient λ_{reg} is fixed at 3.0 in all of our experiments. We optimize Equation (8) by the warm started shooting method [52].

Data Term. For each camera m, we take a 3D point $\mathbf{v}$ on $\mathcal{M}$, whose 2D projection locates on the projected mesh contour. We search a close enough edge pixel $\mathbf{p}$ in the camera m, where its color and edge orientation are best consistent with the 2D projection pixel of $\mathbf{v}$. We minimize the distance between $\mathbf{v}$ and the 3D optical ray from $\mathbf{p}$ passing through the optical center of the camera m. Similar to [3], the data term is defined as,

$$E^{data}(\chi) = \sum_{m} \sum_{\mathbf{v} \in T(m)} \left\| \mathbf{v}(\chi) \times \mathbf{n}^t_m - \mathbf{m}^t_m \right\|_2. \tag{9}$$

Here, $T(m)$ is the set of 3D points, whose 2D projections locate on the projected mesh contour at the camera m, $(\mathbf{n}^t_m, \mathbf{m}^t_m)$ is the Plücker coordinate of the 3D optical ray. We use Canny detector to compute the edge orientation of pixels. Color consistency is computed by euclidean distance in RGB space.

Smoothness Prior. We introduce a novel smoothness motion term,

$$E^{reg}(\chi) = \|\nabla \chi\|_1 + \left\|\nabla_p \chi\right\|_2. \tag{10}$$

Here, the second term is the L_2 difference between χ^t and $\bar{\chi}$, which is a pose predicted by a linear 3rd order auto-regression [3] from previous skeleton poses. The first term is the L_1 norm of $\nabla\chi = \chi^t - \chi^{t-1}$, which enforces sparsity on the temporal gradient of χ. This L_1 norm requires the number of moving joints to be as few as possible, which is intuitively true, even for dramatic and fast motions, since many joints are unchanged at neighboring frames.

5 Experiments

We tested our system on several examples with input videos taken by 2-3 handheld cameras. We used the SONY HDR-CX700 camera, GoPro Hero4 and iPhone to capture our examples. All the raw videos were recorded at $1{,}920 \times 1{,}080$ resolution and the frame rates are 60fps. We downsized them to 960×540 for computational efficiency consideration. We also tried the sequence of subject S4 in HumanEvaI dataset [53] since only the 3D surface mesh model of subject S4 is provided and we only used three color cameras data to perform the motion capture. In our current implementation, the sample based optimization takes about 1.2 minutes for each frame. The local refinement takes a few seconds per frame.

Typical Result. We evaluate our method on various different videos. Some typical results are shown in Fig. 8. Please refer to our supplementary videos, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2017.2693151, to see the recovered skeleton in motion. (a) shows a Tai-Chi coacher captured by three handheld camcorders. The view angle difference between neighboring cameras is from 10 to 30 degrees, with average of 16 degrees. The motion capture algorithm correctly tracks all the 500 frames and only the first frame is used to build the foreground and background model. (b) and (f) show fast flip sequences captured by two handheld mobile phones, whose view angles are separated by only 5 degrees in the whole sequence. (c), (d) and (e) are walking sequences captured by two cameras, with 450, 150

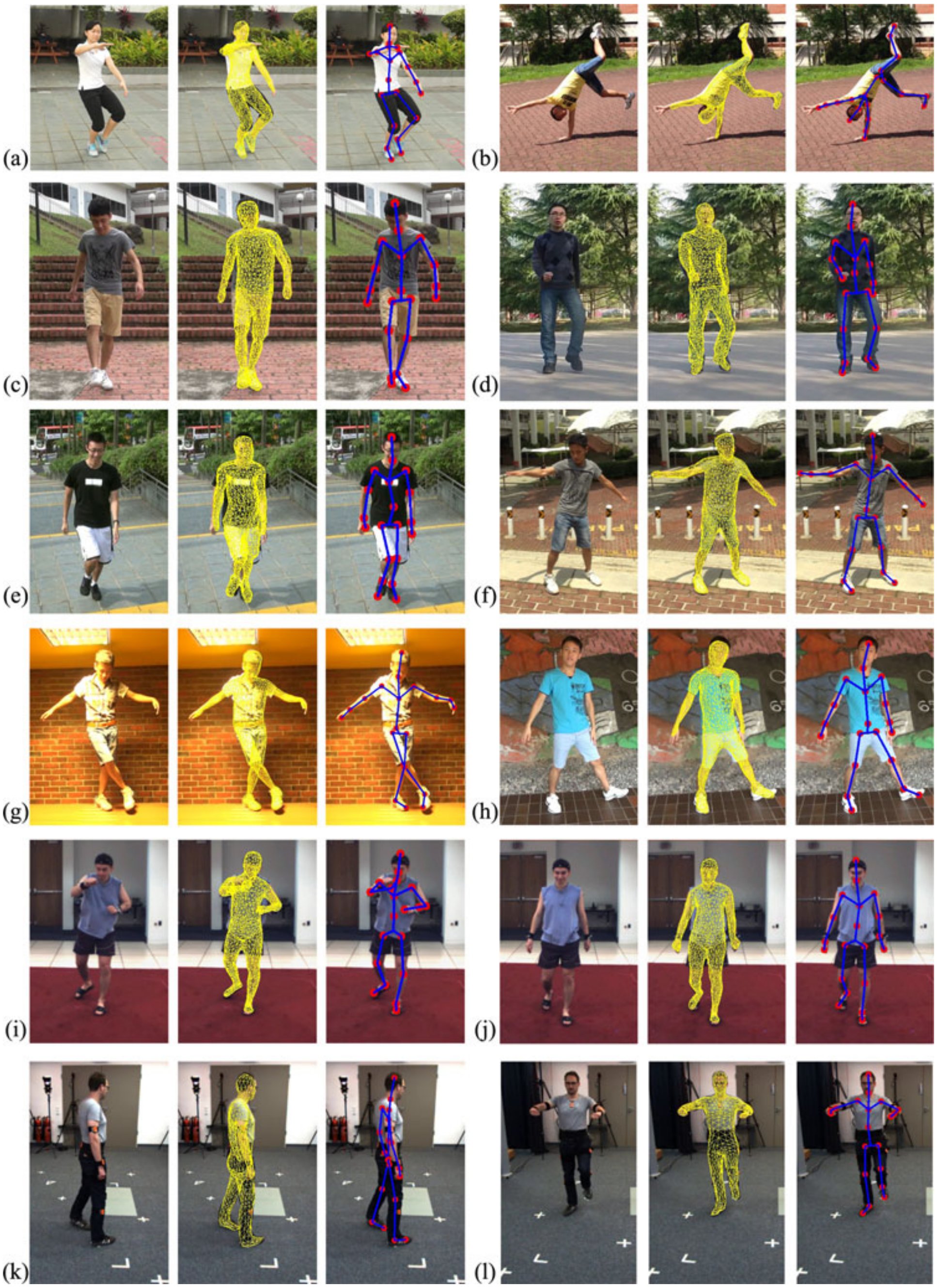


Fig. 8. Typical results including examples with variant illumination, fast motion. The input are 2-3 videos. Our method correctly tracks the whole sequence after building the foreground and background models.

and 200 frames respectively. We use five frames, one frame and three frames respectively to build the appearance model. (g) shows an example captured in a corridor in the evening with two cameras. There are strong illumination changes when the character walks under the ceiling lights. (i) and (j) are the examples from HumanEvaI database. Our method succeeds in all these examples. It demonstrates the strength of our technique that broadens the operation range of markerless motion capture. We also tried the data from TNT16 dataset [54], we utilized three cameras (i.e., camera 0, 2, 4) to perform our method and three frames per 200 frames are selected to build the foreground and background model. Our algorithm succeeds on the video sequence. (k) and (l) in Fig. 8 show the tracking results.

Besides, we captured an example with two handheld cameras in a grove as shown in Fig. 10. There are strong illumination changes when the character walks in the grove. Our method also performs well on this variant illumination case.

For some specific motions, our method can even work with monocular video sequence as shown in Fig. 9. This demonstrates the capabilities of our model-view consistency, which considers both the normalized background and foreground consistency cost in the tracking stage.

Component Evaluation. The importance of several components in the proposed method is evaluated. We have tested the built background color model and the influence of sampling-based global optimization. Besides, we also checked the proposed L_1 regularizer. All the experiments have

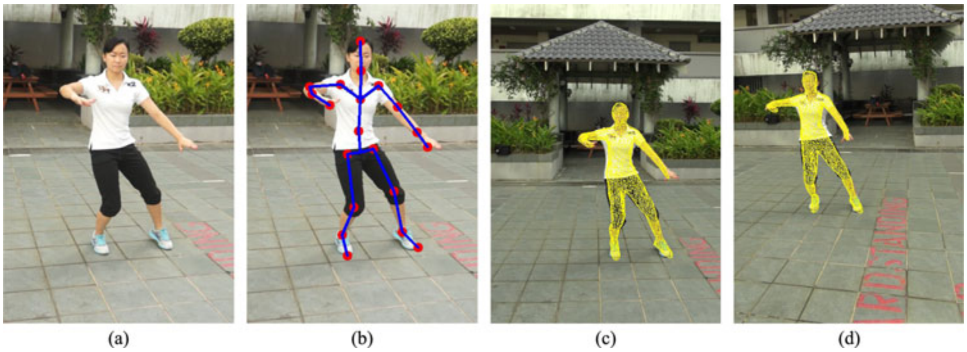


Fig. 9. Monocular tracking result. (a) is the original image, (b) is the tracking skeleton with only one view. We projected the tracked human pose into other two views as shown in (c) and (d).

demonstrated the efficiency and effectiveness of our proposed method.

- *Background Model*. To evaluate the proposed background model, we track the human pose by increasing the number of selected frames to build the background model. The resolution of deformable 2D grid is 16 × 16, which gives 512 triangles. We found that similar results were obtained with the grid resolution from 10 × 10 to 20 × 20. We compared the results of not using background model as well as selecting 1 and 2 frames to build our background model in the tracking stage. The frames are uniformly selected every 150 frames and the number of total frames are 450. Fig. 11 shows the comparison result. This example clearly shows that background model is important for successful tracking and more selected frames to build the background model could improve the robustness of our method.

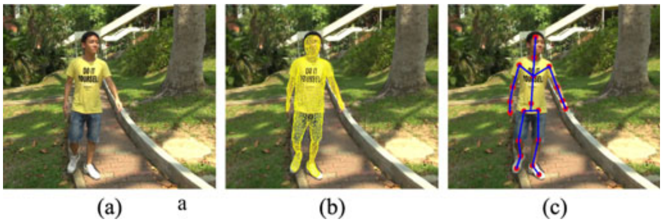


Fig. 10. Variant illumination result. (a) is the original image, (b) and (c) are the projected tracking meshes and skeleton respectively. This example demonstrates that our method also performs well on illumination varying conditions.

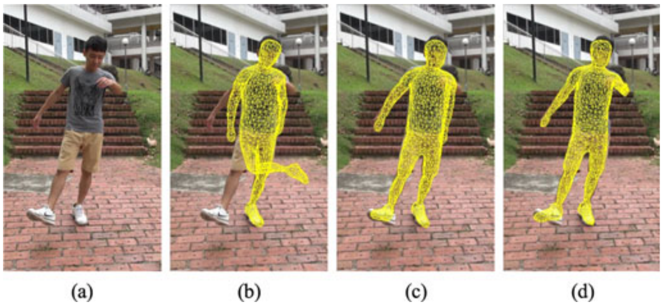


Fig. 11. Comparison with background model evaluation. (a) is the original image, (b) is the tracking result without background model, (c) is the tracking result with one-frame background model, (d) is the tracking result with two-frames background model.

- *Sample-based Pose Optimization*. We dropped the sample-based pose optimization and searched the human pose directly via gradient descendent method to evaluate the importance of sample-based global optimization. Fig. 12 shows the comparison result. It dramatically fails the pose estimation after 16 frames in this example, where there are similar foreground and background colors in the arm region. The reason for tracking failure is from the error accumulation in a frame-by-frame tracking manner and our proposed sample-based pose optimization could alleviate the error accumulation and obtain successful tracking result.
- *Importance of View-View Consistency*. The importance of proposed view-view consistency term is also evaluated in the sample-based global optimization. We dropped the view-view consistency term and perform the sample-based global optimization with only the model-view consistency term in (5). Fig. 13

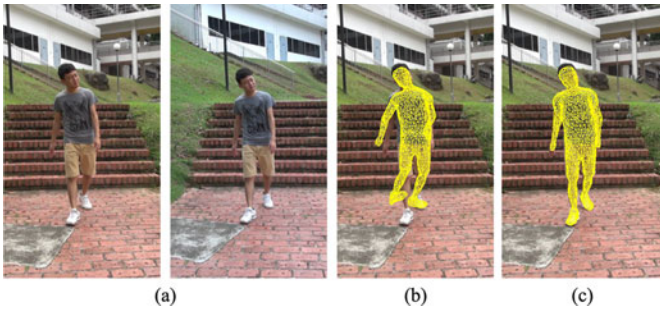


Fig. 12. Comparison with sample-based pose optimization. (a) are two input video frames, (b) and (c) are the tracking results without and with the proposed sample-based pose optimization.

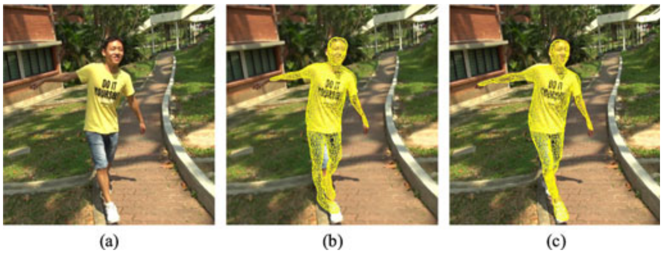


Fig. 13. Comparison with view-view consistency term. 3 cameras which have about 20 degree spanning angles are used to estimate the human pose. (a) is the original image, (b) and (c) are the tracking results without and with the proposed view-view consistency term.

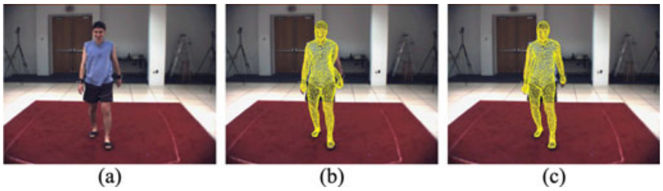


Fig. 14. Comparison with L_1 regularizer. (a) is the original image, (b) and (c) are the estimated poses without and with our L_1-norm based smoothness regularizer. Note the skeleton joint of left arm.

shows the comparison result. It fails when there exist strong shadows since the foreground color model can not model the color distributions well and our proposed view-view consistency term can penalize the ambiguities of illumination variations and obtain successful tracking result.

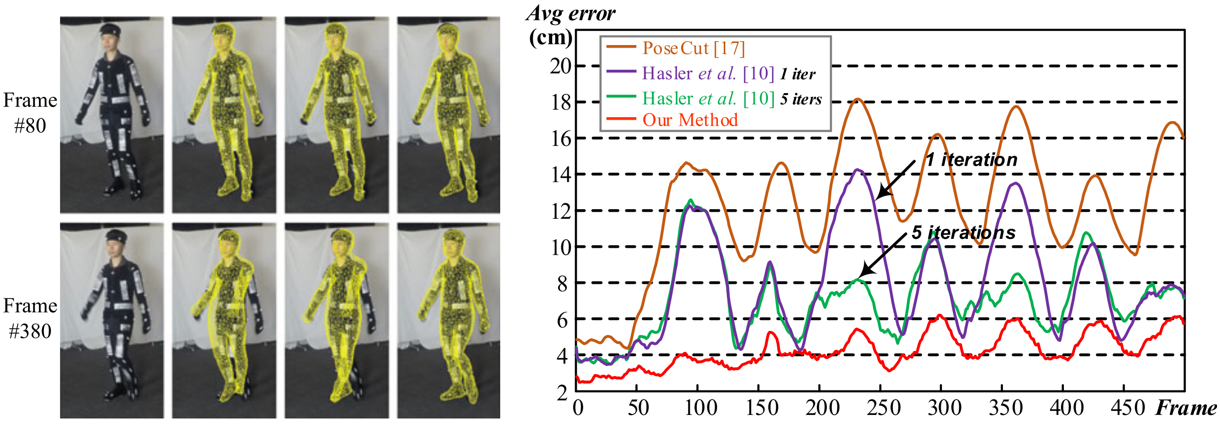
- *Importance of L_1*. Fig. 14 demonstrates the advantage of L_1 regularizer, where (a) are two input video frames, (b) and (c) are the corresponding skeleton poses estimated without and with our L_1-norm based smoothness regularizer. It is clear this regularizer produces better results.

Quantitative Evaluation. To provide a quantitative evaluation, we capture video clips (500 frames) in a Vicon room with two GoPro Hero4 Cameras. We attach some stickers to the black motion capture suite to facilitate feature correspondences. We use only the first frame to build the character and background model. The Vicon system uses a different skeleton model than our system, so we compare the 3D positions of Vicon marker with our tracked mesh model. Specifically, we search the nearest point on the character mesh to each 3D Vicon marker at the first frame and fix the correspondences for the whole sequence. The nearest point is represented as barycentric coordinates on the character mesh. For each frame, we compute the average euclidean distance as the quantitative pose error. The last column of Fig. 15a shows the 3D mesh models recovered by our method. Comparing with this Vicon 'ground truth', the average position error of all marker points varies from 2.48 to 6.20 cm over 500 different frames, with the mean value of 4.23 cm. This error could come from camera calibration, skeleton skinning, and the registration and synchronization of our results with the Vicon 'ground truth'. Considering our simple data capture setup, this error is reasonable. We also tried the method in [4] on our data. It fails on the Vicon room videos because the black motion capture suit is not suitable for its shape-from-shading model. Some qualitative comparisons are included in the supplementary video, available online.

Comparison with Previous Methods. We provide quantitative comparison with PoseCut [19], and the method in [12] on the Vicon sequence. We implemented both methods by ourselves. For the PoseCut algorithm, we adopt the graph-cut library from http://vision.csd.uwo.ca/code/, though dynamic graph-cut is used in the original paper for better efficiency. Since the method in [12] is designed to work with silhouette, we iteratively solved the video segmentation and pose tracking at each frame. We found that the method in [12] reached the local minimum after 5 iterations. We also compared the result after 1 and 5 iterations in this experiment. Due to the simple cameras set up, these two methods can not correctly recover the skeleton motion, as demonstrated in Fig. 15a. We further plot the per frame average position error of Vicon marks in Fig. 15b. PoseCut produces the average error of 12.24 cm and the method in [12] produces 7.06 cm when it converges after 5 iterations.

Limitations and Future Work. Our method relies on the structure-from-motion (SfM) algorithm (CoSLAM [46]) to calibrate camera poses. Thus, it cannot work on examples where the SfM system fails, e.g., at overcrowded scenes. SfM also suffers from drifting errors. In the future, we plan to take a similar strategy as [13] to capture some background images of the static environment beforehand, which can significantly improve the robustness of the SfM system and reduce the drifting error. Our method is not good at dealing with the character with loose clothing since the skinning weights are changed overtime. Our method does not make assumptions about the motions and it could fail for some challenging cases (e.g., a person rotates with the torso). Our method also cannot capture motions of tightly interacting characters.

One possible way to improve our method is to include the discriminative joint detection [15] and put this constraint into the energy minimization. Furthermore, the objective of global optimization and local optimization is not consistent,



(a) (b)

Fig. 15. Quantitative comparison with PoseCut [19] and the method in [12]. (a) shows the result of two selected frames: The 1st column are the original frames, the 2nd and 3rd column are the tracking results via PoseCut [19] and the method in [12] respectively. the 4th column are the results of our method. (b) shows the per frame average Vicon marker position error of different methods.

it would be a future direction to use the recent differential scene model [32] to improve our sample-based global optimization objective and perform the outdoor markerless motion capture with a few hand-held cameras.

The current implementation of our method is time consuming and this time cost mainly comes from the sampling-based global optimization. Future work can be done by several acceleration strategies. We can reduce the time cost by performing the sampling-based optimization in GPU since the data transfer from GPU server to CPU client spends lots of time. We can also modify the global optimization strategies such as 1) performing a motion speed detection and do the global optimization when the motion speed is larger than a threshold; 2) optimizing only some typical DoFs when the differences of the model-view consistency and view-view consistency is larger than a threshold. We hope that these solutions would potentially speed up our algorithm.

6 Conclusions

We propose a method for outdoor markerless motion capture with sparse handheld video cameras. The whole process of the proposed method is simple as frame by frame video segmentation is not required. We advocate a novel model-view consistency formulation that considers both the foreground character and the background. To facilitate the consistency computation of moving background, we model the background as a deformable 2D grid. The character pose is tracked by a global-local optimization. We demonstrate our method on outdoor markerless motion capture with a few hand-held cameras. In the simplest setting, it involves only two mobile phone cameras following the moving character. Our method significantly simplifies the data capture and broadens the operation range of markerless motion capture.

Acknowledgments

We would like thank Kaimo Lin to help us to capture the majority of data. This work was supported in part by the National Key Foundation for Exploring Scientific Instrument No.2013YQ140517, the National Science Foundation of China (NSFC) No.61522111 and No.61531014. Yangang Wang is the corresponding author.

References

[1] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 98.

[2] D. Vlasic, I. Baran, W. Matusik, and J. Popovic, "Articulated mesh animation from multi-view silhouettes," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 97.

[3] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1746–1753.

[4] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt, "On-set performance capture of multiple actors with a stereo camera," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 161.

[5] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 755–762.

[6] J. Shotton, et al., "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1297–1304.

[7] R. Girshick, A. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 415–422.

[8] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *ACM Trans. Graph.*, vol. 31, no. 6, 2012, Art. no. 188.

[9] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real-time human pose tracking from range data," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 738–751.

[10] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon, "Metric regression forests for correspondence estimation," *Int. J. Comput. Vis.*, vol. 113, no. 3, pp. 163–175, 2015.

[11] D. Vlasic, et al., "Practical motion capture in everyday surroundings," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 35.

[12] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel, "Markerless motion capture with unsynchronized moving cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 224–231.

[13] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 31.

[14] G. Pons-Moll, A. Baak, J. Gall, and L. Leal-Taixé, "Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1243–1250.

[15] A. Elhayek, et al., "Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3810–3818.

[16] C.-H. Huang, E. Boyer, N. Navab, and S. Ilic, "Human shape and pose tracking using keyframes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3446–3453.

[17] A. Elhayek, C. Stoll, N. Hasler, K. I. Kim, H. Seidel, and C. Theobalt, "Spatio-temporal motion tracking with unsynchronized cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1870–1877.

[18] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, "Fast articulated motion tracking using a sums of Gaussians body model," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 951–958.

[19] M. Bray, P. Kohli, and P. H. Torr, "PoseCut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 642–655.

[20] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1249–1256.

[21] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3D hand pose estimation from monocular video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1793–1805, Sep. 2011.

[22] M. M. Loper and M. J. Black, "OpenDR: An approximate differentiable renderer," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 154–169.

[23] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 569–577, 2003.

[24] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 126–133.

[25] C. Cagniart, E. Boyer, and S. Ilic, "Free-form mesh tracking: A patch-based approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1339–1346.

[26] P. Huang, M. Tejera, J. P. Collomosse, and A. Hilton, "Hybrid skeletal-surface motion graphs for character animation from 4D performance capture," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 17:1–17:14, 2015.

[27] A. Taneja, L. Ballan, and M. Pollefeys, "Modeling dynamic scenes recorded with freely moving cameras," in *Proc. 10th Asian Conf. Comput. Vis.*, 2011, pp. 613–626.

[28] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1108–1115.

[29] C. Bregler, J. Malik, and K. Pullen, "Twist based acquisition and tracking of animal and human kinematics," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 179–194, 2004.

[30] C.-H. Huang, E. Boyer, and S. Ilic, "Robust human body shape and pose tracking," in *Proc. Int. Conf. 3D Vis.*, 2013, pp. 287–294.

[31] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, "Optimization and filtering for human motion capture," *Int. J. Comput. Vis.*, vol. 87, no. 1/2, pp. 75–92, 2010.

[32] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt, "A versatile scene model with differentiable visibility applied to generative pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 765–773.

[33] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[34] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understanding*, vol. 108, no. 1, pp. 4–18, 2007.

[35] P. Tresadern and I. Reid, "Uncalibrated and unsynchronized human motion capture : A stereo factorization approach," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. I-128–I-134.

[36] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt, "Outdoor human motion capture by simultaneous optimization of pose and camera parameters," *Comput. Graph. Forum*, vol. 34, pp. 86–98, 2014.

[37] X. Wei and J. Chai, "VideoMocap: Modeling physically realistic human motion from monocular video sequences," *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art. no. 42.

[38] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures for multiple human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1669–1676.

[39] H. Joo, et al., "Panoptic studio: A massively multiview system for social motion capture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3334–3342.

[40] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1092–1099.

[41] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 103–110.

[42] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.

[43] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 248.

[44] R. Slyper and J. K. Hodgins, "Action capture with accelerometers," in *Proc. Eurogaphics Symp. Comput. Animation*, 2008, pp. 193–199.

[45] J. Tautges, et al., "Motion reconstruction using sparse accelerometer data," *ACM Trans. Graph.*, vol. 30, no. 3, 2011, Art. no. 18.

[46] D. Zou and P. Tan, "CoSLAM: Collaborative visual SLAM in dynamic environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 354–366, Feb. 2013.

[47] G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," in *Visual Analysis of Humans: Looking at People*. Berlin, Germany: Springer, 2011, ch. 9, pp. 139–170.

[48] C. Wu, "Towards linear-time incremental structure from motion," *IEEE Int. Conf. 3DTV-Conf.*, 2013, pp. 127–134.

[49] I. Baran and J. Popovic, "Automatic rigging and animation of 3d characters," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 72.

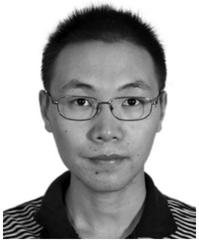
[50] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. Boca Raton, FL, USA: CRC Press, 1994.

[51] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 78.

[52] W. J. Fu, "Penalized regressions: The bridge versus the Lasso," *J. Comput. Graph. Statist.*, vol. 7, no. 3, pp. 397–416, 1998.

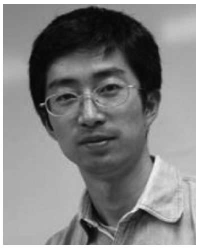
[53] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, no. 1/2, pp. 4–27, 2010.

[54] T. von Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and IMUs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1533–1547, Aug. 2016.



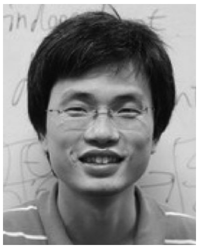
Yangang Wang received the BE degree from Southeast University, Nanjing, China, in 2009 and the PhD degree in control theory and technology from Tsinghua University, Beijing, China, in 2014. He is currently an associate researcher with Microsoft Research Asia. His research interests include image processing, computer vision, computer graphics, and motion capture and animation.

Yebin Liu received the BE degree from the Beijing University of Posts and Telecommunications, China, in 2002 and the PhD degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He was a research fellow in the Computer Graphics Group, Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor with Tsinghua University. His research areas include computer vision, computer graphics, and computational photography.



Xin Tong received the BS and master's degrees in computer science from Zhejiang University, in 1993 and 1996, respectively, and the PhD degree in computer graphics from Tsinghua University, in 1999. He is now a principal researcher in the Internet Graphics Group, Microsoft Research Asia.

Qionghai Dai received the MS and PhD degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively. He is currently a professor in the Department of Automation and the director of the Broadband Networks and Digital Media Laboratory, Tsinghua University, Beijing. He has authored or co-authored more than 200 conference and journal papers and two books. His research interests include computational photography and microscopy, computer vision, and graphics, intelligent signal processing. He is associate editor of the *Journal of Visual Communication and Image Representation*, the *IEEE Transactions on Neural Networks and Learning Systems*, and the *IEEE Transactions on Image Processing*.



Ping Tan received the BS degree in applied mathematics from Shanghai Jiao Tong University, China, in 2000 and the PhD degree in computer science and engineering from Hong Kong University of Science and Technology, in 2007. He joined the Department of Electrical and Computer Engineering with National University of Singapore as an assistant professor in 2007. He received the inaugural MIT TR35@Singapore award in 2012, and the Image and Vision Computing Outstanding Young Researcher Honorable Mention Award in 2012. He is an editorial board member of the *International Journal of Computer Vision*, the *Machine Vision and Applications*. His research interests include computer vision and computer graphics. He is a member of the IEEE and the ACM.