

# Motion Capture from Internet Videos

Junting Dong<sup>1,\*</sup>, Qing Shuai<sup>1,\*</sup>, Yuanqing Zhang<sup>1</sup>, Xian Liu<sup>1</sup>,  
Xiaowei Zhou<sup>1</sup>, and Hujun Bao<sup>2,1,†</sup>

<sup>1</sup> State Key Lab of CAD&CG, Zhejiang University, <sup>2</sup> Zhejiang Lab

**Abstract.** Recent advances in image-based human pose estimation make it possible to capture 3D human motion from a single RGB video. However, the inherent depth ambiguity and self-occlusion in a single view prohibit the recovery of as high-quality motion as multi-view reconstruction. While multi-view videos are not common, the videos of a celebrity performing a specific action are usually abundant on the Internet. Even if these videos were recorded at different time instances, they would encode the same motion characteristics of the person. Therefore, we propose to capture human motion by jointly analyzing these Internet videos instead of using single videos separately. However, this new task poses many new challenges that cannot be addressed by existing methods, as the videos are unsynchronized, the camera viewpoints are unknown, the background scenes are different, and the human motions are not exactly the same among videos. To address these challenges, we propose a novel optimization-based framework and experimentally demonstrate its ability to recover much more precise and detailed motion from multiple videos, compared against monocular motion capture methods.

**Keywords:** Motion capture · Human pose estimation

## 1 Introduction

Human motion capture (MoCap) is a core technology in a variety of applications such as movie production, video game development, sports analysis and interactive entertainment. While there have been some commercial solutions to MoCap, e.g., optical MoCap systems like Vicon, these systems are for professionals but not commodities. The systems are expensive and hard to calibrate. More importantly, the performers need to present in the studio to perform actions, which makes it impossible to collect large-scale motion data for a large population. For example, producing an animated avatar of a celebrity needs to invite the person to the MoCap studio, which is not always feasible especially for amateur productions.

To make human MoCap a commodity, many monocular motion capture algorithms [57, 24, 17, 51] have been developed to recover human motion from single RGB videos. Remarkable progress has been made in past years thanks to the

---

\* Equal contribution. † Corresponding author.

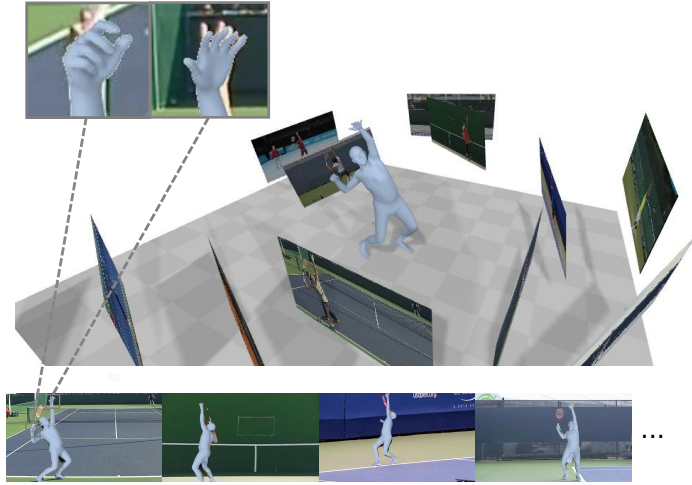


Fig. 1: This paper proposes a system for motion capture from a set of Internet videos which record different instances of the same action of a person. The videos were recorded at different times and in different scenes (bottom). Our system synchronizes the videos, recover the camera viewpoints, and reconstruct the motion accurately (top). More video demonstrations are available at <https://github.com/zju3dv/iMoCap>.

advances in deep learning, public datasets on human bodies, and expressive human models [32, 28]. However, all these methods take a single video as input. As 3D reconstruction from a monocular image is inherently ill-posed, it is extremely difficult to recover accurate and detailed motion from a single video. Leveraging multiple views can resolve the ambiguity, but calibrated and synchronized multi-view videos are not common.

Fortunately, we observe that videos of some celebrities doing some specific actions are abundant on the Internet. While those videos were recorded at different times and the motions in these videos are not exactly the same, they encode the same motion characteristics of the person. Compared to a single video, multiple videos provide richer observations about the specific motion. More importantly, the videos are often recorded at different viewpoints which provide multi-view information to help alleviate the 3D ambiguity and self-occlusion issues.

In this paper, we propose to capture human motion from a collection of Internet videos that record different instances of a person’s specific performance. However, this new problem brings in many challenges that make existing multi-view MoCap algorithms inapplicable: the human motions are not exactly the same among all videos; the videos are unsynchronized; the camera viewpoints are unknown; and the background scenes can be different. To solve these challenges, we propose an optimization-based framework that simultaneously solves video synchronization, camera calibration, and human motion reconstruction. More specifically, the proposed system initializes per-frame 3D human pose estimation with a learned 3D pose estimator, synchronizes videos by matching frames based

on the 3D pose similarity, and jointly optimizes for camera poses and human motions over all the videos. The motions to be recovered are not assumed exactly the same among videos but modeled by a low-rank subspace. Finally, the motion reconstruction and the pose-based video synchronization are iteratively refined. We also show that the video synchronization can be improved by imposing the cycle consistency constraint among multiple videos.

In summary, we make the following contributions:

- We introduce the new task of motion capture from a collection of Internet videos that record different instances of a person’s certain action, which is unexplored in the literature to our knowledge.
- We develop a new optimization-based framework to solve this new task. Our technical contributions include pose-based video synchronization, low-rank modeling of motions, and joint optimization for synchronization, camera poses and human motion.
- We show that, compared to using single videos, the joint analysis of multiple videos provides richer information to address occlusion and depth ambiguity, even if the videos record different motion instances.

## 2 Related work

**Single-view Mocap:** There has been remarkable progress on 3D human pose and shape estimation from single images. Many works focus on the skeleton-based 3D human pose estimation, either first estimating 2D pose from images and then lifting it to 3D [30, 57, 7, 29, 36], or end-to-end regressing to obtain the 3D pose directly [45, 42, 44, 58, 33, 43]. In addition, a lot of works propose to estimate the 3D pose and shape involving a parametric model of the human body [1, 28]. Some early works attempt to use the optimization-based methods [40, 16, 3, 25, 53], which fit the human model to 2D evidence. More recently, many works attempt to directly regress the model from images with a deep network [23, 31, 35, 54, 24, 17, 51]. However, due to the inherent depth ambiguity of single views, the accuracy of these methods is not comparable with the multi-view reconstruction.

**Multi-view Mocap:** Markerless multi-view motion capture has been explored in computer vision for many years. The solutions to this problem are mainly divided into two categories: tracking and pose estimation. Most multi-view tracking methods [15, 2, 26, 27, 11] fit a human body model, e.g., a triangle mesh or a collection of geometric primitives, to image evidence such as keypoints and silhouette. The main difference between them is the type of image evidence and the way to optimize it. However, these tracking based approaches usually require the initialization of the first frame and easily fall into local optima and tracking failures. Hence, more recent works [41, 4, 34, 22] generally tend to estimate 3D human body based on 2D features detected from images. Burenus et al. [4] propose to extend the pictorial structure model to 3D and use it to estimate 3D human skeleton from images. Pavlakos et al. [34] propose to use a ConvNet

for 2D pose estimation and combine with the 3D pictorial structure model to produce 3D pose estimation. Huang et al. [20] and Joo et al. [22] propose to combine statistical body models with a 2D pose estimator and show impressive results. All the above methods assume the multi-view videos are synchronized with known camera parameters.

There are a few methods [18, 12, 13, 55, 49, 52, 38] which attempt to reconstruct the 3D human motion from multiple uncalibrated and unsynchronized videos. Most methods synchronize the videos using additional information, such as audio[18, 12], system time[38], and flashing a light[49], which are unavailable in our scenario. In terms of calibration, many works[18, 12, 55, 49, 52] assume that the camera parameters are provided or obtain the camera geometry using structure from motion based on the static background, which is inapplicable in our setting where the scenes are totally different. Some works [13, 52] also propose to optimize camera parameters and human poses jointly but they assume the motions among videos are exactly the same.

**Video alignment:** When the videos are recording the same event, there are many existing methods to address the temporal alignment problem. Early works [6, 46, 50, 47] generally assume a linear temporal mapping between videos. More recent works propose non-linear solutions based on handcrafted features [48] or learned features [39]. However, for our situation where the videos record similar motions rather than the same event, these approaches are not suitable. Dwibedi et al. [10] propose a self-supervised representation learning method for general video alignment but not tailored for human videos. We will use it as a baseline to evaluate our synchronization component in experiments.

### 3 Methods

Our goal is to reconstruct human motion from multiple videos. Suppose the videos are synchronized, the cameras are calibrated and the motion is the same in these videos, this problem is reduced to a multi-view 3D pose reconstruction problem, which can be solved by first detecting 2D poses in each view and then lifting them to 3D by triangulation. However, this is not the case in our task, where we need to solve video synchronization and motion reconstruction simultaneously with unknown camera geometry, and the motions are similar but not exactly the same across videos.

To solve this challenging problem, we propose an iterative optimization framework that jointly solves synchronization and reconstruction. The intuition is that, if the 3D pose in each video frame is given, we can synchronize videos based on the 3D poses; and if the videos are synchronized, we can recover 3D poses and camera viewpoints from the corresponding frames using multi-view geometry. Figure 2 presents an overview of our approach. We initialize per-frame 3D poses with a CNN-based estimator and iteratively solve synchronization and motion recovery by optimization. In the rest of this section, we first introduce the pose-based video synchronization and then the motion recovery method.



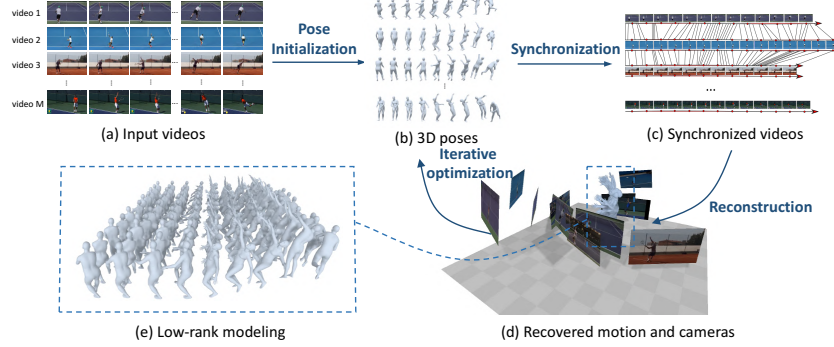


Fig. 2: **Overview of our approach.** Given multiple Internet videos of an action (a), an off-the-shelf 3D human pose estimator is used to initialize the 3D pose of each frame (b). Then, the 3D poses are used to synchronize all videos (c), from which the human motion and camera parameters are recovered (d) with the motion variation across videos modeled by a low-rank matrix (e). Finally, the optimized pose estimates are used to refine the video synchronization again, and the video synchronization and the motion reconstruction are optimized iteratively.

### 3.1 Pose-based video synchronization

In order to leverage multiple views for pose reconstruction, video synchronization is required, i.e., finding the correspondences of frames between videos. However, this is a challenging task because the appearances are very different among videos due to the different background, clothing, and viewpoints. To address this problem, we propose to synchronize videos directly based on 3D human poses seen in the video frames. The initial poses can be obtained by an off-the-shelf pose estimator [24] and refined after synchronization.

Suppose there are  $M$  Internet videos,  $N_j$  is the number of frames for video  $j$ , and  $\mathbf{K}_{ij} \in \mathbb{R}^{3 \times J}$  denotes the 3D human pose estimated for the  $i$ -th frame of video  $j$ . Then, we can measure the likelihood that two frames correspond to each other (a.k.a affinity) based on the similarity between the estimated 3D human poses. Specifically, we compute the Euclidean distance between each pair of 3D poses aligned by the Procrustes method. Then, we map the reciprocal of distance to a value between  $[0, 1]$  as the affinity score between two frames. For a pair of videos  $j_1$  and  $j_2$ , we construct an affinity matrix  $\mathbf{A}_{j_1 j_2} \in \mathbb{R}^{N_{j_1} \times N_{j_2}}$  which consists of all affinity scores between frames of two videos. The correspondences to be estimated can be represented as a partial permutation matrix  $\mathbf{X}_{j_1 j_2} \in \{0, 1\}^{N_{j_1} \times N_{j_2}}$  and efficiently estimated based on  $\mathbf{A}_{j_1 j_2}$  using an optimal assignment algorithm, e.g., dynamic programming considering the sequential constraint on video frames.

If we align each pair of videos separately, the resulting correspondences may be inconsistent due to ignoring the cycle consistency constraint. For example, as shown in Figure 3, the correspondences in green are cycle-consistent since they form a closed cycle and the ones in red are inconsistent. Therefore, we can use



Fig. 3: **An illustration of cycle consistency.** The green lines denote a set of consistent correspondences and the red lines show a set of inconsistent correspondences.

the cycle consistency constraint to improve the alignment of multiple videos. To achieve this, we adopt the result in prior work [19] that the cycle consistency is equivalent to a low-rank constraint on the correspondence matrix  $\mathbf{X}$ , which is the concatenation of all pairwise permutation matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1M} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{M1} & \mathbf{X}_{M2} & \cdots & \mathbf{X}_{MM} \end{pmatrix} \in \mathbb{R}^{N_a \times N_a}. \quad (1)$$

$N_a$  is the number of all frames of all videos.

Therefore, we minimize the following objective function to estimate  $\mathbf{X}$ :

$$f(\mathbf{X}) = \|\mathbf{A} - \mathbf{X}\|_F^2 + \lambda \cdot \text{rank}(\mathbf{X}), \quad (2)$$

where  $\mathbf{A} \in \mathbb{R}^{N_a \times N_a}$  denotes the concatenation of all  $\mathbf{A}_{j_1 j_2}$  similar to the form of  $\mathbf{X}$ ,  $\lambda$  is the weight of low-rank constraint. This problem can be approximately solved with the convex relaxation algorithms in previous work [9, 56]. The relaxed solution  $\mathbf{X}_{j_1 j_2}$  is usually not a valid permutation matrix but a real matrix with values in  $(0, 1)$ , which can be regarded as a denoised version of  $\mathbf{A}_{j_1 j_2}$  with cycle consistency. Finally, to find the frame-to-frame correspondence between video  $i$  and video  $j$ , we use the dynamic time warping algorithm based on the affinity matrix  $\mathbf{X}_{j_1 j_2}$ .

### 3.2 Motion recovery

Even if the videos are synchronized, the problem still cannot be treated as a standard multi-view reconstruction problem for the following two reasons. First, the relative camera poses between videos are unknown and cannot be recovered from structure from motion as the scenes in videos are different. Second, the motions in all videos are not exactly the same. To solve the first issue, we directly register cameras with the human body as the reference and recover the human motion and camera parameters simultaneously. To address the second issue, we

propose to model the motion variation among videos by a low-rank subspace. Before we introduce the methods in detail, we first introduce the representation of human motion.

**Motion representation:** For each video, the corresponding 3D human motion is individually represented by a statistical body mesh model SMPL [28] instead of 3D skeleton, since it contains a richer body prior. The SMPL model is parameterized by the pose parameters  $\theta \in \mathbb{R}^{72}$ , the shape parameters  $\beta \in \mathbb{R}^{10}$ , and a root translation  $\gamma \in \mathbb{R}^3$ , and maps a set of parameters to a body mesh denoted by  $M(\theta, \beta, \gamma) \in \mathbb{R}^{3 \times N_v}$  with  $N_v = 6890$  vertices. A predefined set of 3D body joints  $F(\theta, \beta, \gamma) \in \mathbb{R}^{3 \times J}$  can be generated by linear regression from the mesh vertices, where  $J$  denotes the number of 3D joints. The SMPL+H model[37] which extends SMPL with hands and SMPL-X model [32] which extends SMPL with face and hands can also be used if the video resolution is sufficient for OpenPose[5] to capture the face and hand motion. Our goal is to recover  $\theta_{ij}$ ,  $\beta_{ij}$ , and  $\gamma_{ij}$ , which denote the pose, shape and translation parameters for each frame  $i$  of video  $j$ , respectively. Note that we assume the shape parameters  $\beta_{ij}$  remains the same in one video, i.e.,  $\beta_{ij} = \beta_j$ .

**SMPL-BA:** We attempt to solve camera parameters and SMPL parameters simultaneously by minimizing the reprojection errors of body keypoints detected in video frames, similar to bundle adjustment (BA) in traditional structure from motion. The body keypoints are anchored on the SMPL model. Therefore, we call this procedure SMPL-BA.

Suppose  $R_j^c$  and  $T_j^c$  denote the rotation and translation of the camera  $j$  in the world coordinate system that defines SMPL, respectively. Then, the reprojection error in SMPL-BA can be written as:

$$L_{2d} = \sum_{i,j,z} c_{ijz} \rho \left( W_{ijz} - P \{ R_j^c F(\theta_{ij}, \beta_j, \gamma_{ij})_z + T_j^c \} \right), \quad (3)$$

where  $W_{ijz} \in \mathbb{R}^2$  denotes the  $z$ -th joint of the estimated 2D pose at  $i$ -th frame in video  $j$  with corresponding confidence  $c_{ijz}$  and  $P$  denotes the perspective projection.  $\rho$  denotes the Geman-McClure robust error function for suppressing noisy detection.

In (3), the camera poses  $R_j^c$  and  $T_j^c$  are irrelevant to frame index  $i$ . But in practice the camera may move in each video. To address this issue, we assume that the cameras are only allowed to rotate at fixed camera centers, which is a practical assumption, e.g., in sports broadcasting. Then, we propose to compensate for the camera rotation in each video by warping other frames to the first frame using a homography transformation estimated by feature tracking between frames.

**Low-rank modeling of motions:** When the human motion in each video is not exactly the same, we assume that 3D poses observed in the corresponding frames are very similar which can be approximated by a low-rank matrix:

$$\text{rank}(\theta_i) \leq s, \quad (4)$$

where  $\theta_i = [\theta_{i1}^T; \theta_{i2}^T; \dots; \theta_{iM}^T] \in \mathbb{R}^{M \times 72}$  denotes the collection of pose parameters in all videos of frame  $i$  and the constant  $s$  controls the degree of similarity. Note that each video has its own SMPL parameters. The only constraint that links all videos is the low-rank constraint, which is soft and allows difference among videos.

In addition, we also assume that the 3D trajectories of the root joint of the body should be similar among videos. Suppose the root trajectories in all videos are denoted by  $\gamma = [\gamma_1^T; \gamma_2^T; \dots; \gamma_M^T] \in \mathbb{R}^{M \times 3N}$ , where  $\gamma_j \in \mathbb{R}^{3N}$  is the trajectory in video  $j$  and  $N$  is the number of frames. Then, the constraint can be written as:

$$\text{rank}(\gamma) \leq s. \quad (5)$$

Intuitively, how much motion variation is allowed depends on the selected rank  $s$ . A larger  $s$  allows larger variation but imposes less constraint across views. We empirically find that  $s = 1$  or  $2$  works well, which imposes sufficient multi-view constraint while allowing reasonable motion variation across videos.

**Objective function:** Combining all discussed above, the final objective function to optimize can be written as:

$$\begin{aligned} \min \quad & L_{2d} + \lambda_t L_{temp}, \\ \text{s.t.} \quad & \text{rank}(\theta_i) \leq s, i = 1, 2, \dots, N, \\ & \text{rank}(\gamma) \leq s, \end{aligned} \quad (6)$$

where  $L_{temp}$  is a temporal smoothing term with weight  $\lambda_t$  to eliminate jittering in motion:

$$L_{temp} = \sum_{i=1}^{N-1} \|\theta_i - \theta_{i+1}\|_F^2. \quad (7)$$

**Optimization:** To simplify the optimization, we introduce two auxiliary variables  $\mathbf{Z}_i \in \mathbb{R}^{M \times 72}$  and  $\mathbf{Y} \in \mathbb{R}^{M \times 3N}$  to decouple the rank constraints with the objective function:

$$\begin{aligned} \min \quad & L_{2d} + \lambda_t L_{temp} + \lambda_{r_1} \sum_{i=1}^N \|\theta_i - \mathbf{Z}_i\|_F^2 + \lambda_{r_2} \|\gamma - \mathbf{Y}\|_F^2, \\ \text{s.t.} \quad & \text{rank}(\mathbf{Z}_i) \leq s, i = 1, 2, \dots, N, \\ & \text{rank}(\mathbf{Y}) \leq s \end{aligned} \quad (8)$$

where  $\lambda_{r_1}$  and  $\lambda_{r_2}$  are weighting parameters.

The problem in (8) is highly nonconvex. However, reliable initialization allows us to use local optimization to solve this problem. Specifically, we update each variable alternately while the others remain fixed. The pose  $\theta_i$ , shape  $\beta_j$ , and translation  $\gamma$  parameters of SMPL can be updated with Gradient Descent. It is a standard low-rank approximation problem to update  $\mathbf{Z}_i$  and  $\mathbf{Y}$ , which can be solved by SVD analytically. The update of  $\mathbf{R}_j^c$  and  $\mathbf{T}_j^c$  can be solved with a perspective-n-point (PnP) algorithm that minimizes reprojection errors over all frames of video  $j$ .

**Initialization:** We initialize the SMPL parameters for each frame using a pre-trained neural network [24], which is further refined by minimizing the reprojection error of 2D keypoints for each frame. Next, the videos are initially synchronized based on the initial pose estimates as introduced in Section 3.1. Then, a reference video is selected, whose camera coordinate system is regarded as the world frame. Note that the initial SMPL model in each video is defined in the coordinate system of the respective camera. Therefore, the relative camera poses between two videos can be initialized by rigidly aligning the SMPL models, assuming the SMPL pose parameters are the same between videos. When intrinsics are unknown, we set the focal length to be a large constant, approximating a weak-perspective camera model. In this way, the camera poses can be initialized.

### 3.3 Iterative optimization

The video synchronization in the first iteration may not be very accurate based on initial pose estimates. Therefore, we propose to refine the synchronization based on the optimized poses. More specifically, the affinity matrix  $\mathbf{A}$  in Section 3.1 is updated with the optimized poses given by the SMPL-BA and the frame correspondences are re-computed using the new affinity matrix. Then, the SMPL-BA is computed again with the updated synchronization. Both synchronization and reconstruction benefit from each other in iterative optimization, which will be experimentally demonstrated in Section 4.2.

## 4 Experiments

### 4.1 Motion Capture from Internet videos

There is no existing dataset for our task. Therefore, we collect a new dataset that consists of 20 actions of various actors, such as tennis serves, yoga and Tai Chi. Take tennis serves as an example. We download the publicly available videos of some tennis players from YouTube, and manually crop the videos roughly to obtain a set of video clips of serves for each player. Figure 4 shows the statistics of the number of videos and average number of frames for each action. The dataset is available at <https://github.com/zju3dv/iMoCap>.

We apply the proposed approach on each action of this dataset to recover the corresponding human motion. Some representative results are visualized in Figure 6, which shows that the proposed approach is able to recover 3D human motion as well as camera geometry from these videos, even if they were recorded at different times. These videos record the action from very different viewpoints and therefore provide multi-view constraints to help alleviate the depth ambiguity and self-occlusion issues that often occur for single-view estimation. Consequently, compared to the monocular motion capture algorithm [24], our approach produces much more detailed and faithful motion, as indicated by the circles in Figure 6. In addition, with the multi-view constraint, our method is also able to recover an accurate 3D trajectory of the body as shown

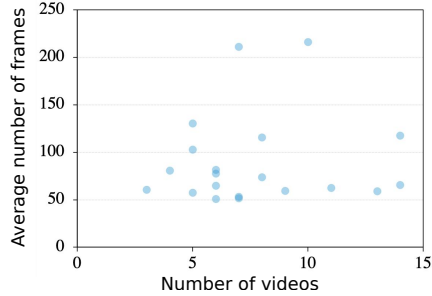


Fig. 4: **Collected Internet video dataset.** Each point denotes one action.

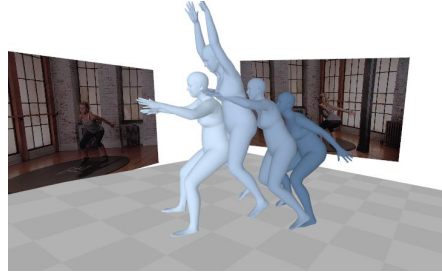


Fig. 5: **Trajectory recovery.** Our approach is able to recover the absolute 3D trajectory of human motion. The brightness of human mesh indicates the chronological order.

in Figure 5, which is infeasible for monocular motion capture algorithms. Note that, the proposed approach can be easily extended to hand motion recovery if the 2D hand pose estimation is available as shown in Figure 1 and 6 (Tai Chi). We find that most of the failure cases are because of failed 2D pose estimation. Also, when the viewpoints of videos are similar, the depth ambiguity cannot be resolved even if multiple videos are used. *More qualitative results and video demonstrations are available in the supplementary material.*

Since the motion in all the videos is not exactly the same, each video has its own SMPL parameters with a low-rank constraint to make the parameters correlated among multiple videos. An alternative is to assume the motions are all the same and use a single model with the same set of SMPL parameters for all videos. We provide a qualitative comparison in Figure 7. More specifically, we reproject the initial 3D mesh, the 3D mesh reconstructed by our low-rank model, and the 3D mesh reconstructed by the single model to images and compare them in terms of 2D consistency. The results show that the projected mesh using the single model is less consistent with 2D evidence as shown in the red circles, which suggests that the single model cannot model the motion difference among videos. Low-rank modeling is able to capture more detailed motion, such as the curved back of the performer, as shown in the green circles. Overall, low-rank modeling recovers more detailed and natural motion than the results of a single model and initial monocular results.

## 4.2 Quantitative evaluation

While we have collected a dataset of Internet videos to demonstrate the qualitative performance of our system, *quantitative* evaluation is difficult due to the lack of 3D ground truth, a similar case for most prior work on reconstruction from Internet data. For quantitative analysis, we synthesize a dataset using existing datasets [21] with ground-truth annotations. We select some challenging sequences in the Human3.6M dataset [21] and modify the data to simulate the

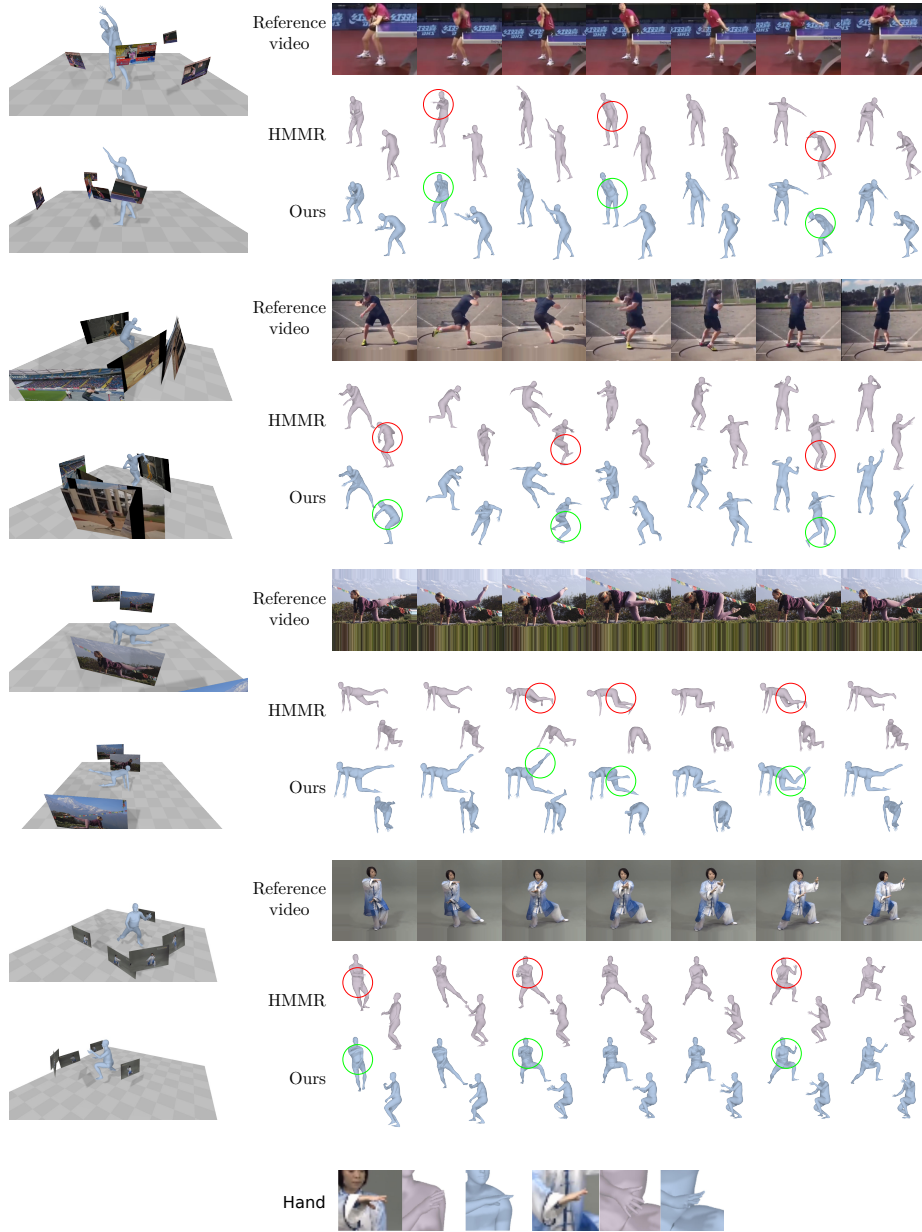


Fig. 6: **Results on Internet videos** of table tennis serves, shotput, yoga, and Tai Chi (with hands motion). The left images present the reconstructed human motion and camera positions visualized in two viewpoints. On the right, we present some frames of the reference video and corresponding motion capture results from HMMR [24] and our method. Red and green cycles emphasize some representative differences between the results from HMMR and our method.

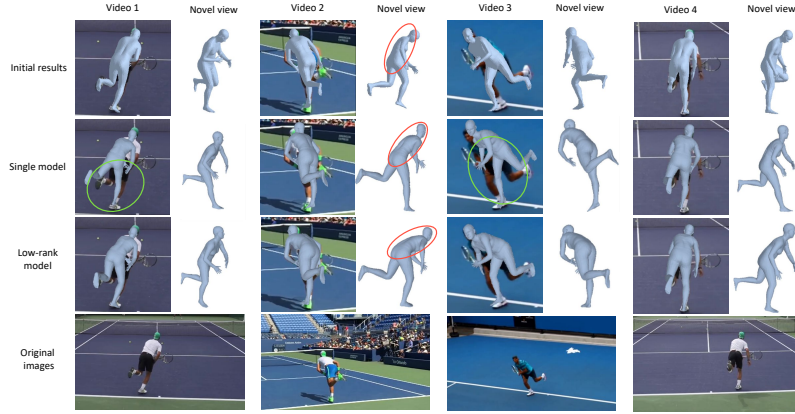


Fig. 7: **Effect of low-rank modeling.** The projected mesh of a single model is less consistent with 2D evidence (red marks). A low-rank model is able to capture differences among videos and recover more accurate motion such as the curved back of the performer (green marks).

unsynchronized and uncalibrated scenario. Please refer to Figure 8 for details. We would like to note that the purpose of evaluation on Human3.6M is not to compare against existing methods in the standard Human3.6M setting, but to provide an ablative analysis of our system when solving the proposed problem.

**Video synchronization:** As described in section 3.1, we propose a pose-based video synchronization method to address the different appearances among videos, i.e., background, clothing, and viewpoints. We also impose the cycle consistency constraint to improve the synchronization. Here, we compare with some baselines and use the standard video alignment metric to measure the alignment of two videos. In particular, for each frame of non-reference video  $v_i$ , we compute the frame distance between the matched frame and the ground truth position in reference video  $v_0$  and normalize it by the video clip length.

We first propose a simple alternative that uses the DP algorithm to quantize the original affinity matrix directly. The result of this baseline method (‘No cycle-consis’) is shown in Table 1. The results show that imposing the cycle consistency constraint can reduce the alignment error of video synchronization significantly.

Another baseline is a recent self-supervised representation learning method [10] based on the cycle consistency loss to align videos. Their network is retrained on the evaluated sequences and the alignments are obtained by the dynamic time warping algorithm on the features. The results of their method (‘TCC’) on the dataset are presented in Table 1. The results show that our 3D pose based method outperforms the generic method by a large margin in our problem.



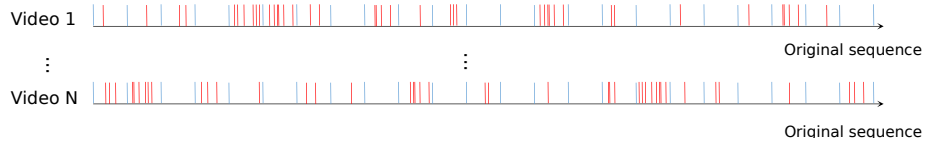


Fig. 8: **Dataset generation for quantitative analysis.** We edit the videos in Human3.6M to simulate the unsynchronized scenario. As the dataset is large, we only select a few actions, i.e., SittingDown, Sitting, Smoking, Photo and Phoning. For each action, we first sample  $N_{s1}$  frames at equal intervals (**blue lines**) from each video, which results in  $N_{s1} - 1$  segments. Then we randomly choose  $N_{s2}$  segments and randomly sample  $N_{s3}$  frames (**red lines**) from each selected segment. In our experiments, we set  $N_{s1} = 150, N_{s2} = 50$  and for each segment  $N_{s3}$  is a variable value randomly selected from 1 to the length of the segment. The dataset is available at <https://github.com/zju3dv/iMoCap>.

Table 1: **Quantitative analysis of synchronization.** ‘No cycle-consis’ denotes our synchronization method without cycle consistency constraint. ‘TCC’ represents the general video synchronization method [10] based on representation learning.

| Method          | Synchronization error |
|-----------------|-----------------------|
| Ours            | <b>0.77%</b>          |
| No cycle-consis | 1.19%                 |
| TCC [10]        | 11.24%                |

**Reconstruction:** We evaluate the motion reconstruction quantitatively. To evaluate 3D joint error, we use the standard metric, i.e., the mean per joint position error (MPJPE) and the error after rigid alignment (P-MPJPE).

As videos are unsynchronized and uncalibrated, none of the existing multi-view MoCap methods is applicable to the proposed problem. Monocular MoCap methods are the only applicable alternatives. We compare with the state-of-the-art monocular method HMMR [24] and the results are shown in Table 2. Our method significantly reduces the reconstruction error compared to HMMR, which shows the benefit of using multiple videos.

Note that we use a generic 2D pose detector [14] which is not fine-tuned on Human3.6M. With no fine-tuning, we wish to evaluate the generalization ability of our system when applied to unseen and challenging videos. We also report the results with a fine-tuned 2D pose detector[8] in Table 2, which show that using a fine-tuned detector significantly reduces the reconstruction error.

In addition, we validate the influence of the number of videos. We report the reconstruction accuracy with various numbers of videos in Table 2. The results show that more videos improve the accuracy of reconstructed motion. As for Internet videos, while more views generally improve the results, we empirically find that three or four videos are sufficient in most cases.

Table 2: **Quantitative analysis of reconstruction.** ‘HMMR’ denotes the state-of-the-art monocular motion capture method [24].

| Method                       | MPJPE (mm)   | P-MPJPE (mm) |
|------------------------------|--------------|--------------|
| HMMR                         | 109.80       | 78.26        |
| Ours+generic 2D, 4 videos    | 76.48        | 53.34        |
| Ours+fine-tuned 2D, 1 video  | 80.65        | 62.58        |
| Ours+fine-tuned 2D, 2 videos | 78.45        | 59.42        |
| Ours+fine-tuned 2D, 3 videos | 71.48        | 53.77        |
| Ours+fine-tuned 2D, 4 videos | <b>66.53</b> | <b>50.33</b> |

Table 3: **Quantitative analysis of iterative optimization.** ‘No iter-opt’ denotes our method without iterative optimization of synchronization and reconstruction.

| Method      | Synchronization error | MPJPE (mm)   | P-MPJPE (mm) |
|-------------|-----------------------|--------------|--------------|
| Ours        | <b>0.77%</b>          | <b>66.53</b> | <b>50.33</b> |
| No iter-opt | 0.91%                 | 71.49        | 54.12        |

**Iterative optimization:** Our approach iteratively optimizes video synchronization and reconstruction to let them benefit from each other. ‘No iter-opt’ in Table 3 indicates the result without such an iterative optimization, which shows that iterative optimization reduces both alignment and reconstruction errors.

## 5 Summary

In this paper, we demonstrated the potential of leveraging multiple Internet videos to recover accurate and detailed human motion, which in a long-term perspective opens up the possibility of collecting high-quality and diverse human motion data for free from existing Internet videos. Unlike standard multi-view motion capture, in this new task the human motions are not exactly the same among all videos; the videos are unsynchronized; the camera viewpoints are unknown; and the background scenes can be different. All these challenges make existing multi-view motion capture algorithms inapplicable. To address all challenges above, we proposed (1) low-rank modeling of motions to handle motion variation among videos; (2) pose-based multi-video synchronization and calibration; and (3) *most importantly* a unified optimization-based framework to solve the entire problem, which doesn’t treat synchronization, calibration and motion recovery as separate tasks, but integrates them in a single optimization problem. Both qualitative and quantitative results demonstrated the effectiveness of the proposed approach. Please see the supplementary material for more video demonstrations.

**Acknowledgement:** The authors would like to acknowledge support from NSFC (No. 61806176) and Fundamental Research Funds for the Central Universities (2019QNA5022).

## References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM transactions on graphics (TOG) (2005)
2. Bo, L., Sminchisescu, C.: Twin gaussian processes for structured prediction. IJCV (2010)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016)
4. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: CVPR (2013)
5. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Real-time multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
6. Caspi, Y., Irani, M.: Spatio-temporal alignment of sequences. T-PAMI (2002)
7. Chen, C.H., Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In: CVPR (2017)
8. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded Pyramid Network for Multi-Person Pose Estimation (2018)
9. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: CVPR (2019)
10. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: CVPR (2019)
11. Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In: CVPR (2015)
12. Elhayek, A., Stoll, C., Hasler, N., Kim, K.I., Seidel, H.P., Theobalt, C.: Spatio-temporal motion tracking with unsynchronized cameras. In: CVPR (2012)
13. Elhayek, A., Stoll, C., Kim, K.I., Theobalt, C.: Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In: Computer Graphics Forum (2015)
14. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)
15. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. IJCV (2010)
16. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV (2009)
17. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019)
18. Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: CVPR (2009)
19. Huang, Q.X., Guibas, L.: Consistent shape maps via semidefinite programming. In: Computer Graphics Forum (2013)
20. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 3DV (2017)
21. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. T-PAMI (2013)

22. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
23. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
24. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
25. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR (2017)
26. Lee, C.S., Elgammal, A.: Coupled visual and kinematic manifold models for tracking. IJCV (2010)
27. Li, R., Tian, T.P., Sclaroff, S., Yang, M.H.: 3d human motion tracking with a coordinated mixture of factor analyzers. IJCV (2010)
28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) (2015)
29. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
30. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: CVPR (2017)
31. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 3DV (2018)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
33. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: CVPR (2018)
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: CVPR (2017)
35. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: CVPR (2018)
36. Pavlo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
37. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) (2017)
38. Saini, N., Price, E., Tallamraju, R., Enficiaud, R., Ludwig, R., Martinovic, I., Ahmad, A., Black, M.J.: Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In: ICCV (2019)
39. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: ICRA (2018)
40. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NeurIPS (2008)
41. Sigal, L., Isard, M., Haussecker, H., Black, M.J.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. IJCV (2012)
42. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017)
43. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)

44. Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: ICCV (2017)
45. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. CVPR (2017)
46. Tuytelaars, T., Van Gool, L.: Synchronizing video sequences. In: CVPR (2004)
47. Ukrainitz, Y., Irani, M.: Aligning sequences and actions by maximizing space-time correlations. In: ECCV (2006)
48. Wang, O., Schroers, C., Zimmer, H., Gross, M., Sorkine-Hornung, A.: Videosnapping: Interactive synchronization of multiple videos. ACM Transactions on Graphics (TOG) (2014)
49. Wang, Y., Liu, Y., Tong, X., Dai, Q., Tan, P.: Outdoor markerless motion capture with sparse handheld video cameras. TVCG (2017)
50. Wolf, L., Zomet, A.: Wide baseline matching between unsynchronized video sequences. IJCV (2006)
51. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)
52. Xu, X., Dunn, E.: Discrete laplace operator estimation for dynamic 3d reconstruction. arXiv preprint arXiv:1908.11044 (2019)
53. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR (2018)
54. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3d sensing of multiple people in natural images. In: NeurIPS (2018)
55. Zheng, E., Ji, D., Dunn, E., Frahm, J.M.: Sparse dynamic 3d reconstruction from unsynchronized videos. In: ICCV (2015)
56. Zhou, X., Zhu, M., Daniilidis, K.: Multi-image matching via fast alternating minimization. In: ICCV (2015)
57. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: CVPR (2016)
58. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)