Fully Automatic Multi-person Human Motion Capture for VR Applications

Ahmed Elhayek[1,2], Onorina Kovalenko[1], Pramod Murthy[1,2], Jameel Malik[1,2], and Didier Stricker[1,2]

[1] German Research Centre for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[2] University of Kaiserslautern, Germany
{ahmed.elhayek, onorina.kovalenko, pramod.murthy,
jameel.malik, didier.stricker}@dfki.de

Abstract. Fully automatic tracking of articulated motion in real-time with monocular RGB camera is a challenging problem which is essential for many virtual reality (VR) applications. In this paper, we propose a novel temporally stable solution for this problem which can be directly employed in VR practical applications. Our algorithm automatically estimates the number of persons in the scene, generates their corresponding person specific 3D skeletons, and estimates their initial 3D locations. For every frame, it fits each 3D skeleton to the corresponding 2D body-parts locations which are estimated with one of the existing CNN-based 2D pose estimation methods. The 3D pose of every person is estimated by maximizing an objective function that combines a skeleton fitting term with motion and pose priors. Our algorithm detects persons who enter or leave the scene, and dynamically generates or deletes their 3D skeletons. This makes our algorithm the first monocular RGB method usable in real-time applications such as dynamically including multiple persons in a virtual environment using the camera of the VR-headset. We show that our algorithm is applicable for tracking multiple persons in outdoor scenes, community videos and low quality videos captured with mobile-phone cameras.

Keywords: Human motion capture · Convolutional neural network · anthropometric data.

1 Introduction

Human motion capture has applications in many fields such as VR, augmented reality (AR), 3D character animation (i.e. for movies and games), human-computer interaction, and sports. The last decade have witnessed significant progress in marker-less human motion capture approaches which work directly on real-world video streams [39, 44, 49]. Although, many marker-less algorithms have achieved high accuracy under challenging conditions, most commercial VR systems still use marker-based algorithms that require to place markers on the human body. One of the main reasons is that marker-less algorithms require several manual initialization steps (e.g. 3D human model generation and initial pose estimation) which are cumbersome, require a lot of experience and time consuming.

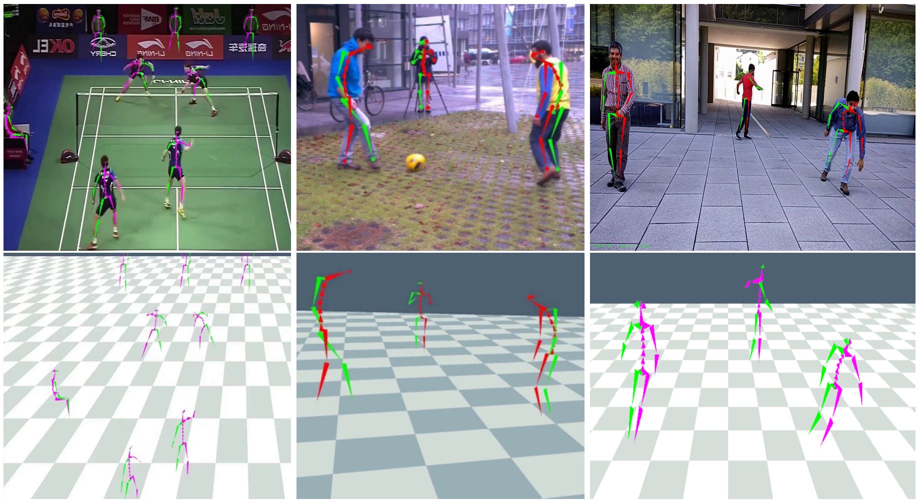


Fig. 1: Our algorithm recovers 3D skeletons poses in real-time. It captures complex motions of 8 persons in a community video (left), 3 persons in a video from the Marconi [19] datasets (middle) and 3 persons in a video captured with our mobile-phone RGB camera (right). Top row shows overlaid 2D skeletons and bottom row shows 3D visualizations of the captured skeletons.

Monocular RGB cameras are very common in many VR-headsets, laptops, and smartphones. Thus, developing a fully automatic real-time multi-person marker-less human motion capture algorithm that works with such monocular cameras is essential for many VR applications. An example of these applications is to include and animate multiple 3D characters in a VR environment using the camera of a VR-headset. Furthermore, this algorithm allows to interface PCs, laptops, or smartphones with their cameras (e.g. play games). However, developing such algorithm is challenging and requires 1) automatic estimation of number of persons in the scene 2) automatic generation of their 3D skeletons 3) automatic estimation of their initial 3D location 4) dynamical generation or deletion of 3D skeletons for persons entering or leaving the scene; respectively 5) real-time multi-person fitting energy function.

Most of marker-less approaches estimate the articulated joint angles of moving subjects from multi-view video recordings [51, 19, 21, 22]. These algorithms require manual estimation of persons number, their 3D models, and their initial poses. Moreover, they fail to reliably track articulated motion in general scenes with single RGB camera. While many recent algorithms have managed to estimate accurate human motion from monocular depth cameras [5, 57, 16], only few algorithms work accurately with monocular RGB cameras [38, 58, 37]. Although some of these algorithms achieve better accuracy than our algorithm, they do not succeed under our challenging multi-person tracking conditions. For instance, [38] does not succeed with multi-person and assumes an initial human pose to be given. Moreover, it's skeleton initialization requires given 2D body parts detections from several frames and height of the person. In addition to these limitations, other monocular algorithms such as [58, 37] are offline and exhibits jitter over time due to per frame estimation. To the best of our knowledge, our algorithm is the first that performs automatic personalized skeleton generation and initial pose localization of varying number of persons in real-time. Moreover, it reconstructs the motion of multi-person in real-time using a single off-the-shelf RGB camera.

Our algorithm allows to overcome the limitations of RGB-D cameras which fail in general outdoor scenes due to sunlight interference. These cameras have lower resolution, limited range, higher power consumption, and are not widely available as RGB cameras. Our algorithm is able to track multiple persons moving in front of cluttered and non-static backgrounds with moving low quality camera which suffers from high distortion. It also succeeds in case of strong illumination changes. It works with any mobile-phone cameras, webcams, and community videos (e.g. YouTube videos). Our novel algorithmic contributions that enable this, are:

1. Real-time, simple and automatic multi-person human 3D skeletons generation; see Section 4.1.
2. Automatic initial 3D location estimation of each person in the scene; see Section 4.2.
3. Automatic detection of the change in number of persons and generating or deleting the corresponding 3D skeletons on the fly while tracking; see Section 4.3.
4. Novel algorithm which tracks full articulated joint angles of multiple persons at high accuracy and temporal stability in real-time, given 2D body-part locations; see Section 4.3.

The estimated multi-person motions can be used in many fields such as VR, AR, motion-driven 3D game character control, and human computer interaction. Furthermore, our algorithm can be optimized for smartphones and driving assistance applications. In our experiments, we show that our algorithm can capture even complex and fast body motion of multi-person in real-time; see Figure 1. We managed to capture complex motions of multiple persons in outdoor scenes with a moving mobile phone camera, a spherical camera in a car, and a webcam in an office.

2 Related Work

Video-based human motion capture has seen great advances in recent years. We refer the reader to the surveys [39, 44, 49] for an overview. We focus the discussion in this section on two categories: methods based on multi-view input and methods that rely on a monocular RGB camera.

Multi-view: Most multi-view marker-less motion capture setups employ a human 3D model whose pose parameters are computed by optimizing an overlap measure between the projected 3D model and the input images. They attain high accuracy by tracking the human model over the image sequence with offline computation [10, 9, 50]. In [24], the pose is estimated from silhouette and color information. The approaches presented in [7, 30, 33] use training data to learn a motion model or a mapping from image features to the 3D pose. Tracking without silhouette information is also possible by combining model-guided segmentation and pose estimation. Earlier methods, such as [43], attempted to capture human skeletal motion from stereo footage, but did not achieve the same accuracy as methods using dense camera setups.

Amin et al. [3] propose a multi-view pictorial structures model that incorporates evidence across multiple viewpoints to allow robust 3D pose estimation. Belagiannis et al. [6] extend [3] for 3D pose estimation of multiple humans. However, a common problem with these approaches is jitter due to missing temporal information at each time step. The approach by [51] introduced an analytic formulation for calculating the model-to-image similarity based on a Sums-of-Gaussians model. Other works extend multi-view motion capture approaches towards tracking with moving or unsynchronized cameras [25, 48, 21, 22]. These methods need separate initialization (e.g. using [8, 46] at the beginning of each sequence and after loss of track in local minima of their non-convex fitting functions). Robustness can be increased with a combination of generative and discriminative estimation [19, 45]. An accurate manually initialized human 3D model is essential for these methods. We propose an approach for automatic multiple skeletons generation which avoids using human model projection to speed up estimation. This allows to utilize generative tracking components and ensure temporal stability.

Monocular RGB: Depth-based motion capture methods [57, 16] have achieved robust real-time results. However, in this section, we focus on RGB-based methods. These methods can be divided into generative and discriminative methods. The generative motion capture problem is fundamentally under-constrained in case of monocular input. Thus, it is only successful for motion capture from short clips and when combined with strong motion priors [54]. Manual annotation and correction of frames is suitable for some applications such as actor reshaping in movies [28] and garment replacement in videos [47]. These generative algorithms preclude live applications because of manual interaction and expensive optimization.

Recently, many monocular discriminative human pose estimation methods have been introduced. Some of them discriminatively learned mapping from the image directly to human joint locations [1, 29, 27]. CNN based 2D and 3D human pose estimation approaches achieve state-of-the-art accuracy. For instance, [34, 36, 52, 17] estimate human 3D pose directly from monocular image or video. Chen et al. [15] automatically synthesize training images with ground truth pose annotations and train CNNs with these synthetic images for 3D pose estimation.

Other approaches estimate 3D human pose from 2D body parts locations in a monocular image [31, 2, 55, 32, 23]. Many of these works have been realized by assuming manually labeled 2D body part locations. Recently, many CNN-based 2D pose estimation methods were proposed [13, 26, 11, 56, 53, 14]. All these methods provide 2D body parts locations which can be used for 3D human pose estimation. For example, Cao et al. [13] managed to efficiently detect the 2D poses of multiple persons in an image using a nonparametric representation, which allows to learn associations between body parts of each individual in the image. Bogo et al. [8] used 2D body parts locations detected by [42] to automatically estimate the 3D pose and shape of the human body from a single unconstrained image. However, this method is not real-time and works for single person only.

Most closely related to the present paper are approaches for real-time recovery of 3D human pose with monocular RGB camera. Only a few methods target this problem for temporally stable results which is directly usable in practical applications. The top

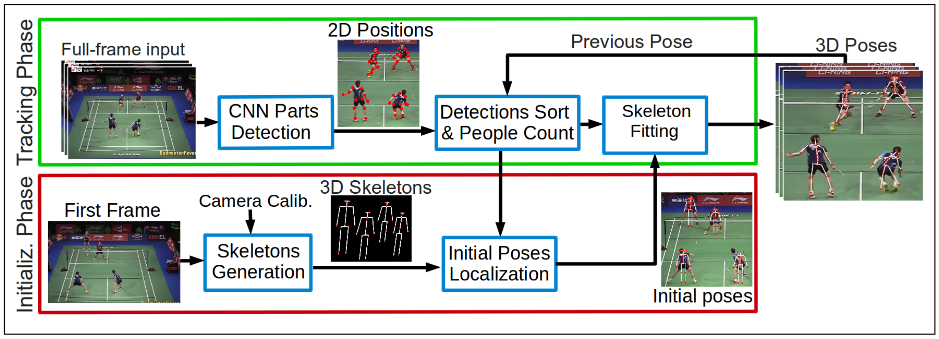


Fig. 2: Overview. We generate multiple person-specific 3D skeletons based on anthropometric data, and estimate the initial location of each person in an initialization phase (bottom, Section 4.1). In the tracking phase, we estimate 2D body-parts positions from the input video streams. These 2D positions are used to estimate global 3D poses by skeleton fitting (top, Section 4.3). The *Dynamic Scene Update* step generates or deletes 3D skeletons for persons who enter or leave the scene.

performing single RGB 3D pose estimation methods are based on CNNs [38, 58, 41, 37, 35]. Mehta et al. [37] use a 100-layer CNN architecture to predict 2D and 3D joint positions simultaneously. However, [37] is unsuitable for real-time execution due to the additional preprocessing steps such as bounding box extraction. Mehta et al. [38] propose a 3D pose estimation approach that uses CNN to detect 2D and 3D pose jointly. Then, an optimization based skeletal fitting method is applied to estimate 3D poses in real-time. All these methods, however, work for single person only. On the other hand, we propose a multi-person 3D pose estimation approach which automatically estimates person-specific 3D skeleton and initial 3D location for each person in the scene. Thereafter, the pose of every person is estimated by means of optimizing an energy function for multi-person skeleton fitting.

3 Overview

Input to our approach can be either the live stream of a monocular RGB camera (e.g. webcam or VR-headset), YouTube video, or video captured with a mobile-phone camera. Any of these inputs yield a single frame I_i at discrete points in time $i = \{1, 2, 3, ...\}$. For frame I_i, the final output is $\mathbf{X} = \{X_1, ..., X_{prsn}\}$ where $prsn$ is the number of persons in the scene . X_j is the 3D skeletal pose parameters of the person with index j. This output is temporally consistent and in global 3D space which makes it perfect for applications such as virtual reality and character control. Our algorithm works with any camera (i.e. moving, static, webcam, or spherical camera with strong distortion) and general scenes (i.e. indoors or outdoors with strong illumination changes).

An outline of the processing pipeline is given in Figure 2. Many human motion capture algorithms such as [20, 21, 51] assume given person-specific 3D skeletons and initial pose parameters X_{init}. This number of skeletons is fixed over the whole sequence. In contrast to these algorithms, we automatically estimate the number of persons in the scene. Then, we automatically generate person-specific 3D skeletons and estimate the initial location of each person in the scene. All these automatic steps are done in real-time at the beginning of each sequence which we refer to as **initialization phase**. The basic idea of our automatic skeleton generation approach is to adapt a default human skeleton to the length of each bone of each person. To this end, anthropometric data tables are used to define the length of each bone as a function of the height of each person; see Section 4.2 for details.

Given the person-specific 3D skeletons, it is still not possible to start the tracking process without defining the initial pose of each person. Existing human motion capture algorithms either estimate the initial pose manually or use computationally expensive methods such as [8]. In this paper, we automatically estimate the 3D root location of each person in the scene which resolves this limitation; see Section 4.2 for details.

In the tracking phase, we start with a CNN-based approach [13, 11] to estimate the 2D locations of the body-parts for each person in the scene. The output of this step is the matrix $\mathbf{J} = [J_1, ..., J_{prsn}]$ where J_i contains body-parts locations of person i. However, the order and number of the persons in $\mathbf{J}$ may vary from frame to frame. Therefore, we use Equation 4 to find the 2D body-parts positions J_i corresponding to specific 3D skeleton. Thereafter, we dynamically generate 3D skeletons for persons who enter the scene and delete the skeletons of those who left; see Section 4.3 for details.

The pose parameters $\mathbf{X} = \{X_1, ..., X_{prsn}\}$ are optimized given the 2D body-parts positions with the following energy function at each time frame I_i:

$$E(\mathbf{X}, \mathbf{J}) = E_{FIT}(\mathbf{X}, \mathbf{J}) - w_L E_L(\mathbf{X}) - w_A E_A(\mathbf{X}) \tag{1}$$

where $E_{FIT}(\mathbf{X}, \mathbf{J})$ is the skeletons fitting term (Section 4.3). $E_L(\mathbf{X})$ enforces joint limits, and $E_A(\mathbf{X})$ is a smoothness term penalizing strong accelerations; see [51] for details. The weights $w_l = 0.1$ and $w_a = 0.05$ were found experimentally and are kept constant in all experiments. This energy function is smooth and analytically differentiable. Thus, it can be optimized efficiently using standard gradient ascent initialized with the initial pose estimated in Section 4.2.

4 Real-time Multi-person 3D Human Pose Estimation

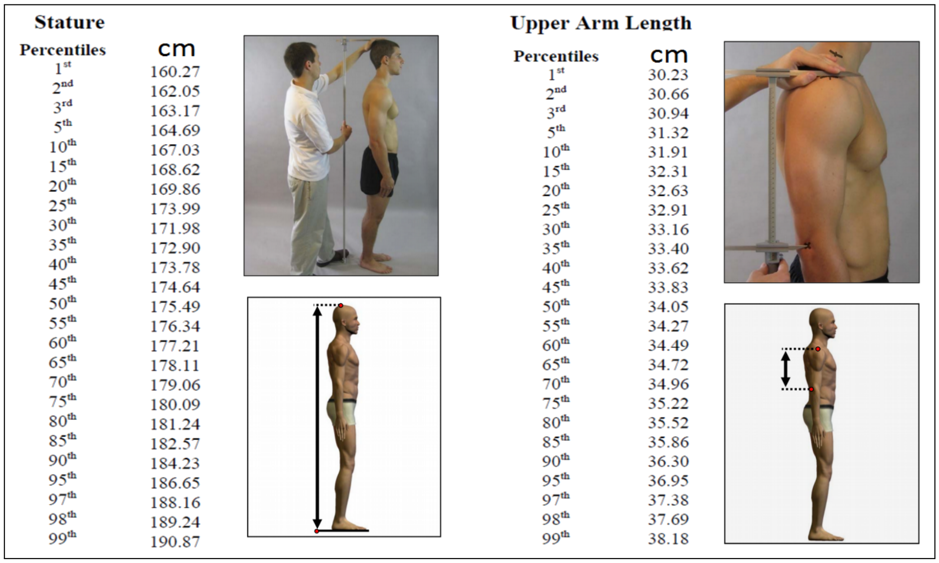
In this section, we describe in detail the components of our fully automatic algorithm which captures articulated skeleton motion of several subjects in general scenes from monocular RGB input. The initialization phase is discussed in Section 4.1 and Section 4.2, while the tracking phase is explained in Section 4.3.

4.1 Automatic 3D Skeletons Generation

Human motion capture algorithms require human 3D model with properly personalized skeleton and/or body shape and appearance to successfully track a single person. Many algorithms consider model personalization as a different problem and use manual or semi-automatic model generation approach, which greatly reduces their applicability. In this section, we propose a novel automatic approach that generates a skeleton specific to each person.

In [46], an automatic algorithm that jointly creates skeleton and body model of a single person is presented. However, this algorithm requires many RGB cameras to estimate the body model. In [19, 21], the skeleton and the body model of each person is generated in a semi-automatic way from a set of calibration poses prior to motion recording. Nonetheless, in case of no control over the footage and person motion, their method fails. Therefore, developing a simple, efficient, and automatic human 3D skeleton estimation approach is very important as it enables our solution to be adopted in more practical applications where the manual model generation is not feasible. We propose the first skeleton generation approach to automatically estimate skeletons for many persons in real-time.

In our approach, we generate a default skeleton for every person. The initial number of persons is automatically estimated given the 2D detections of the first frame. Then, we adapt the bone length of each skeleton to match the corresponding person. Our default skeleton consists of 25 bones and 26 joints. Each joint is defined by an offset to its parent joint and a rotation represented in axis-angle form. In total, the model consists of 73 parameters (70 rotational and 3 translational); see [19] for details. The anthropomorphic data tables [12] allow to define the length of each bone in the skeleton as a function of the height of the person. Figure 3 shows part of the anthropomorphic data table which defines the relation between the length of the upper arm bone and the height of the person. With these tables, the skeleton generation task is simplified to the estimation of a single parameter (i.e. the height of the person). Inspired by [40, 17], the height of each person can be estimated from monocular RGB camera by back-projecting 2D features of an object into the 3D scene space. The output of this step is a person-specific human 3D skeleton for every person in the scene.



Stature

Percentiles	cm
1st	160.27
2nd	162.05
3rd	163.17
5th	164.69
10th	167.03
15th	168.62
20th	169.86
25th	173.99
30th	171.98
35th	172.90
40th	173.78
45th	174.64
50th	175.49
55th	176.34
60th	177.21
65th	178.11
70th	179.06
75th	180.09
80th	181.24
85th	182.57
90th	184.23
95th	186.65
97th	188.16
98th	189.24
99th	190.87

Upper Arm Length

Percentiles	cm
1st	30.23
2nd	30.66
3rd	30.94
5th	31.32
10th	31.91
15th	32.31
20th	32.63
25th	32.91
30th	33.16
35th	33.40
40th	33.62
45th	33.83
50th	34.05
55th	34.27
60th	34.49
65th	34.72
70th	34.96
75th	35.22
80th	35.52
85th	35.86
90th	36.30
95th	36.95
97th	37.38
98th	37.69
99th	38.18

Fig. 3: Part of the anthropometric data tables which is used for person-specific 3D human skeletons generation: height data table (left), the corresponding table of upper arm length [12] (right).

4.2 Multi-person Skeleton Localization

Given the personalized skeleton, the motion capture process can not start without initial 3D pose of each person. This essential initialization is, unfortunately, neglected by many methods and solved with manual initialization step, or with a different computationally expensive approach such as [8]. As our algorithm is stable even with inaccurate initial poses, we simplify the initial pose estimation problem to the estimation of the initial root position (i.e. 3D point between hips) of each person. To this end, we use the heights H_i^{3D} of each person i, their 2D body-part detections in the first frame J_i , and the monocular camera focal length f. The individual heights H_i^{3D} can be estimated as in Section 4.1, while the 2D body-parts detections J_i are estimated using the CNN-based algorithm; see Section 4.3 for details. As the upper body is usually more visible than the lower body, we use the height of the torso $H_{trs,i}^{3D} \approx 0.3 * H_i^{3D}$ for estimating the root depth. The 2D height of the torso $H_{trs,i}^{2D}$ is the distance between the neck $j_{nck,j}$ and the root $j_{rt,i} = (j_{lhip,i} + j_{rhip,i})/2$. With this, the depth of the root is calculated by:

$$z_i^{3D} = \frac{H_{trs,i}^{3D} * f}{H_{trs,i}^{2D}}. \tag{2}$$

Then, the 3D root position is calculated by:

$$\{x_i^{3D}, y_i^{3D}, z_i^{3D}\} = \mathbf{\Phi}^{-1}(j_{rt,i}^{x} * z_i^{3D}, j_{rt,i}^{y} * z_i^{3D}, z_i^{3D}) \tag{3}$$

where $\mathbf{\Phi}$ is the projection operator. Thereafter, each skeleton is automatically moved such that its root position matches the root location of the corresponding person in 3D space.

4.3 Skeleton Fitting for Dynamic Number of Persons

In the initialization phase, personalized skeletons and their initial 3D locations are estimated in real-time once at the beginning of the tracking process. On the other hand, the tracking phase is repeated for every frame. The first step of the tracking phase is the estimation of the 2D body-parts positions. Recently, many CNN based methods managed to accurately estimate these 2D body-parts positions [13, 26, 11]. Although, any of these methods can be used in our framework, we used both [13] and [11] in our experiments. As [13] achieves state-of-the art accuracy with multi-person, the majority of our results are based on this algorithm. Therefore, in this section, we assume, without loss of generality, that 2D body-part positions are estimated with [13].

The 2D body-part detection algorithm does not have any temporal relation between consecutive frames. Thus, the order of the resulting 2D body-part detections in $\mathbf{J} = [J_1, ..., J_{prsn}]$ for one frame can be different the previous frame. This means that the body-parts positions J_m may correspond to a different person in each frame. For this reason, the next step in our tracking phase is to associate each existing 3D skeleton with the corresponding 2D detections J_m in each frame. To this end, we define a similarity measure between the skeleton defined by pose parameters X_k and $J_m = [j_{m,1}, ...j_{m,prt}]$ where prt is the number of 2D body part detections of one person. This is done by first projecting the 3D joint positions defined by X_k into the 2D image plane using the projection operator Φ. Thereafter, the distance between each projected 3D joint and the corresponding 2D detection is calculated. The final similarity between skeleton with index k and detections in J_m is defined as follows:

$$SIM_{k,m} = \sum_{l=1}^{n_{prt}} \|\mathbf{\Phi}(\mathbf{f}_{k,l}(X_k)) - j_{m,l}\| \tag{4}$$

where $\mathbf{f}_{k,l}$ is the 3D joint position corresponding to the 2D body part $j_{m,l}$. At the end of this step, each skeleton with index k will be associated with the 2D detection J_i where $i = \arg\min_x SIM_{k,x}$.

For tracking varying number of persons, we need to generate a new 3D skeleton for each person who enters the scene and remove the skeleton of those who leave the scene. After associating each 3D skeleton with the corresponding 2D detections J_i, some items of $\mathbf{J}$ may be left without a corresponding 3D skeleton. These items correspond to either persons who just entered the scene or false positive detection of a human. To distinguish between these two cases, we use the confidence of each body part detection in J_i which is an additional output of the CNN-based approach. This confidence allows to compute a score for each J_i which corresponds to probability of a new person entering the scene. For each new J_i with score above the threshold $\alpha = 0.5$, we generate 3D skeleton for the corresponding person and estimate the respective initial 3D location. On the other hand, in case of a person leaving the scene or largely occluded, J_i corresponding to an existing skeleton will either have very low score or disappear from $\mathbf{J}$. In both cases, we remove that skeleton.

Our multi-person skeleton fitting term measures the similarities between a given skeleton pose X_n corresponding to one of the persons and 2D body-parts positions J_n of that person. Similar to Equation 4, we project each 3D joint position and calculate the distance to the corresponding 2D detection $j_{n,l}$. The final fitting term is defined as:

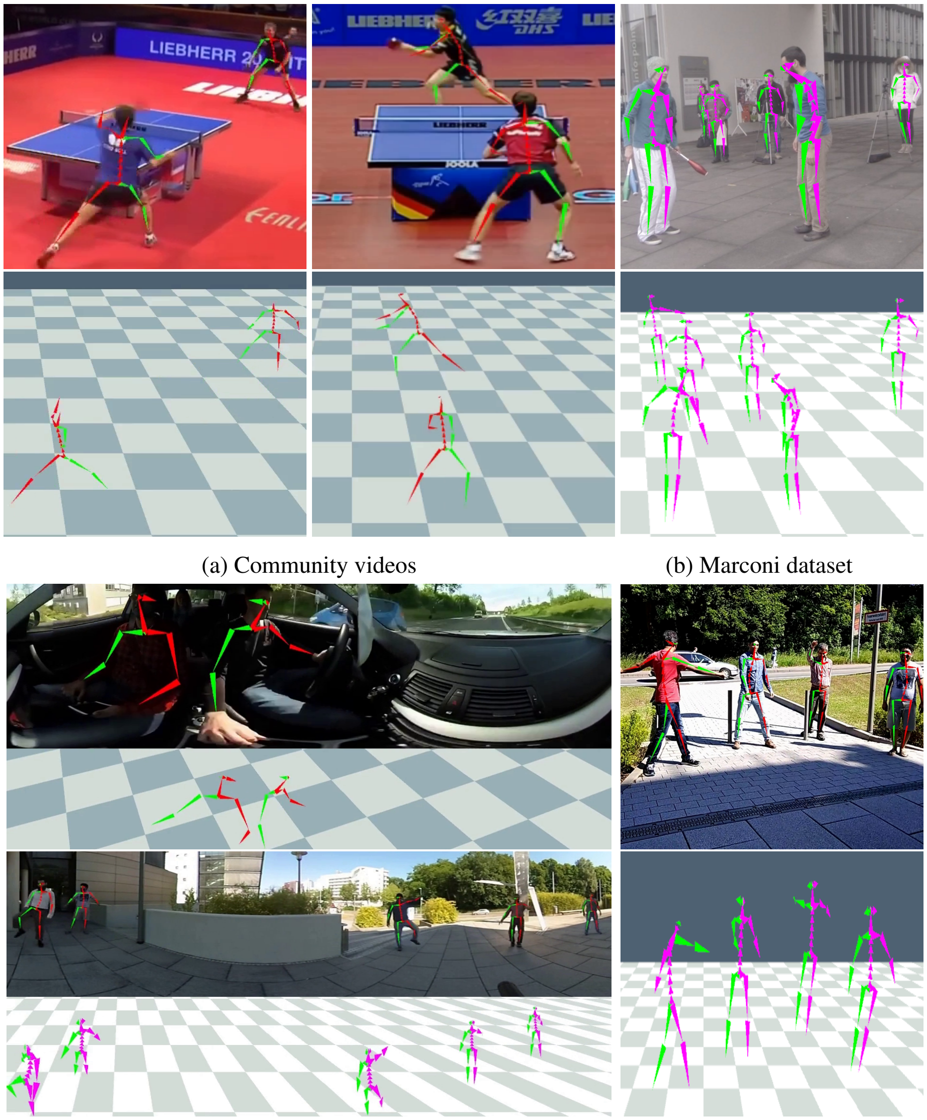
$$\begin{gathered} E_{FIT}(X,J) = \\ \sum_{n=1}^{n_{prsn}} \sum_{l=1}^{n_{prt}} w(j_{n,l}) \exp\left(-\frac{\|\mathbf{\Phi}(\mathbf{f}_{n,l}(X_n)) - j_{n,l}\|^2}{\sigma^2} \right) \end{gathered} \tag{5}$$

where $w(j_{n,l})$ is the confidence of the 2D body-parts detection $j_{n,l}$. This confidence is estimated by the CNN body-parts estimation method.

Applying per-frame pose estimation techniques on a video does not ensure temporal consistency of motion. Thus, small pose inaccuracies lead to temporal jitter. Therefore, we combine our multi-person skeletons fitting energy with temporal filtering and smoothing in a joint optimization framework to obtain an accurate, temporally stable and robust result; see Equation 1.

5 Experiments and Results

We demonstrate the effectiveness of our algorithm through experimental evaluations of more than 20 challenging real world sequences. Some of these sequences were acquired from community videos including varying number of persons performing complex and



(a) Community videos

(b) Marconi dataset

(c) Spherical camera images

(d) Mobile phone camera

Fig. 4: Sample results with overlaid 2D skeletons estimated with **Implementation 1** (top) and respective 3D reconstructions (bottom) which show successful multi-person tracking in challenging scenarios. (a) shows multi-person pose results over YouTube videos playing table tennis and fencing sports. (b) shows results over selected difficult sequences from Marconi dataset. (c) shows pose estimation results inside a car and outdoor scene recorded using a spherical RGB camera. (d) shows tracking results with strong illumination changes in outdoor scene captured using mobile phone camera

fast motions. We also captured many outdoor and indoor sequences with mobile-phone and spherical camera. One of the outdoor sequences was recorded in car with spherical camera to illustrate the usefulness of our algorithm for applications such as driving assistance system. We performed live tracking of multiple persons at around $23Hz$ with low quality webcam. In addition to that, we used many sequences from the Human3.6M [27] and the Marconi [19] datasets. These sequences vary in numbers and identities of persons, complexity and speed of the motion, the lighting conditions, cameras types (e.g. mobile-phone, GoPro, spherical cameras, and webcams), the frame resolutions, and the frame rates. Our algorithm is the first multi-person monocular human motion capture method which does not require any manual work for 3D human model and initial pose adaptation. It automatically generates 3D skeletons and estimates initial poses for multiple person. It operates with input images without the need of bounding box cropping. As a result of this, our experimental setup is very simple. Given the input images and the focal length of a single RGB camera, we produce high quality reconstruction results. Qualitative results can be viewed in accompanying supplementary video. The run-time of our algorithm depends on the number of persons in the scene, the complexity of the motion and the resolution of the input frames. Our computations are performed on a 8-core Xeon CPU and a GeForce GTX 1080 GPU. Although our algorithm's implementation is not yet well optimized for improved run-time performance, average processing time of a single frame from a single person sequence (e.g. the Greeting sequence from the Human3.6M dataset [27]) is 44 milliseconds. The 2D body parts detection [13] takes 32 milliseconds while the 3D skeleton fitting takes 12 milliseconds. Given the body parts detections of the first frame and the height of each person, the initialization phase takes around 0.01 milliseconds.

Our algorithm is not restricted to use a particular 2D body-parts detection method. Hence, we show results of our algorithm with two different body parts detection methods. The first implementation **Implementation 1** uses [13] for 2D body-parts detections. This implementation is discussed in details in Section 4. Notably, in contrast to other 2D body part detection methods, [13] does not require cropping to track multi-person sequences. On the other hand, our second implementation **Implementation 2**, which is based on [11], requires cropping of every person. However, our algorithm can perform cropping automatically and without significant change to our original pipeline in Figure 2. To this end, the rough pose of each person is estimated by extrapolating his pose from the previous frame. The bounding box of each person is estimated by projecting each 3D skeleton to the camera view. This allows to crop and scale each person. With this additional automatic step, [11] can be used instead of [13] in our pipeline for 2D body part detections.

Qualitative Results: We used our first implementation **Implementation 1** to track mroe than 15 sequences. Sample frames from the tracked sequences are shown in Figure 1 and Figure 4. Please, see the supplementary video for more detailed tracking results. Our algorithm successfully estimated the pose parameters of multiple persons in challenging outdoor and indoor sequences with monocular RGB camera. This shows the ability of our algorithm to successfully track sequences with many (i.e. up to eight) persons performing complex and fast motions under strong lighting variations and strong distortion. Previous monocular methods such as [38, 58, 37] fail to track these sequences

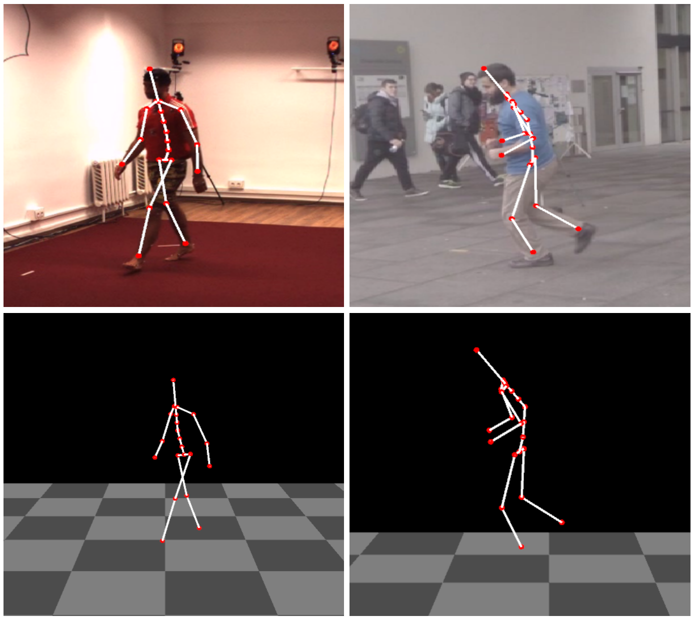


Fig. 5: Sample images from the H3.6M dataset (left column) and the Marconi dataset (right column) with overlaid 2D Skeleton along-with respective 3D pose recovery using **Implementation 2** .

in real-time. We also tracked a sequence captured in car and several sequences captured with mobile-phone. This shows that our approach is suitable for practical applications in different fields including VR. In Figure 5, we show the 3D pose reconstruction results based on our second implementation **Implementation 2**. Two sequences from the public datasets the Human3.6M and the Marconi are successfully tracked.

To demonstrate the usefulness of our algorithm for real-time applications (e.g. dynamically including multiple persons in a virtual environment using the camera of the VR-headset), we tracked the motion of multiple persons from live stream of webcam. Figure 6 shows that our real-time 3D pose estimation provides a natural motion interface in challenging scenarios. Furthermore, we capture sequence with a mobile-phone camera where several people enter and leave the scene. Our algorithm succeed in automatically detecting the change in number of persons and generating or deleting the corresponding 3D skeletons on the fly while tracking; see the supplementary video.

Comparison: In Figure 7, we compare the accuracy of our algorithm with the accuracy of [38, 18] on two challenging sequences. Our algorithm managed to accurately track all the persons in two sequences; see the supplementary video for more detailed tracking results. While [18] work only offline, [38] achieved lower tracking accuracy for only one of the two persons in the scene.

System Components Evaluation: We quantitatively evaluate the importance of the components of our algorithm by creating different alternatives of it. The first alternative is constructed by removing the skeleton generation step. This means that the default

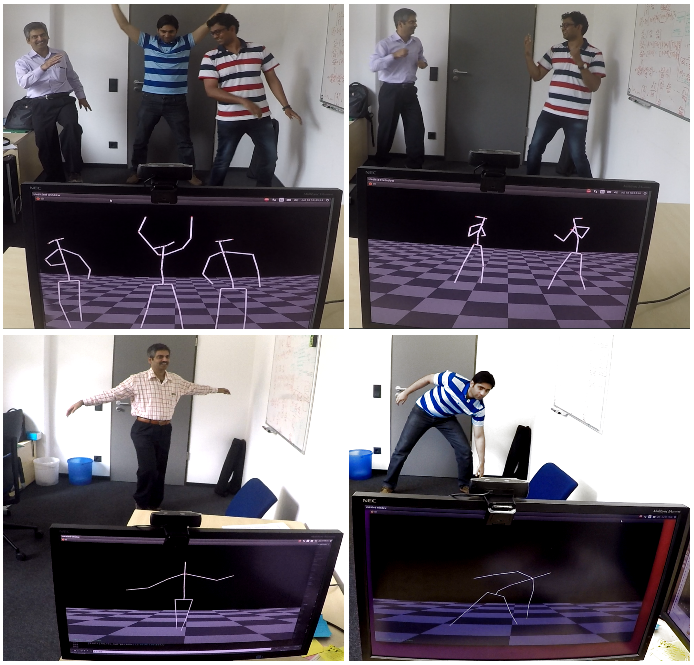


Fig. 6: The real-time 3D pose estimation with **Implementation 1** (Top) and **Implementation 2** (Bottom). Our algorithm provides a natural motion interface on images from live webcam video.

skeleton is used without adaptation to the tracked person. The second alternative is constructed by removing the initial pose localization step where the initial pose parameters are set to zero or to random values. We evaluated these alternatives by tracking the *Walking* sequence from Human3.6M dataset [27] which captures Subject $S9$. The Mean Per Joint Position Error (MPJPE) with our complete algorithm is 90mm while it is 460mm without the first alternative. The second alternative fails completely because the energy function is non-convex which leads to stuck in a local maxima; see Figure 9 and the supplementary video.

Quantitative Evaluation: We quantitatively evaluate our algorithm using the *Directions*, *Posing* and *Waiting* sequences from Human3.6M dataset [27] which capture Subject S9. Figure 8 shows sample images with overlaid 2D skeletons and respective 3D reconstructions from these sequences. The average error of all frames of these three sequences is $159.33mm$. [38] achieves lower error with monocular RGB camera. However, the CNN body-parts detector of [38] is trained on images from the test dataset (i.e. the Human3.6M dataset [27]). On the other hand, the CNN body-parts detectors which we use, are trained on different datasets such as the MPII Human Pose dataset [4].

Discussion: Our approach is subject to a few limitations. Currently, the depth estimation of our algorithm is not very accurate, especially in case of occlusion of wrists and ankles. This causes relatively higher 3D joint position errors in comparison to other methods. However, this is also a common problem with approaches relying on a monoc-

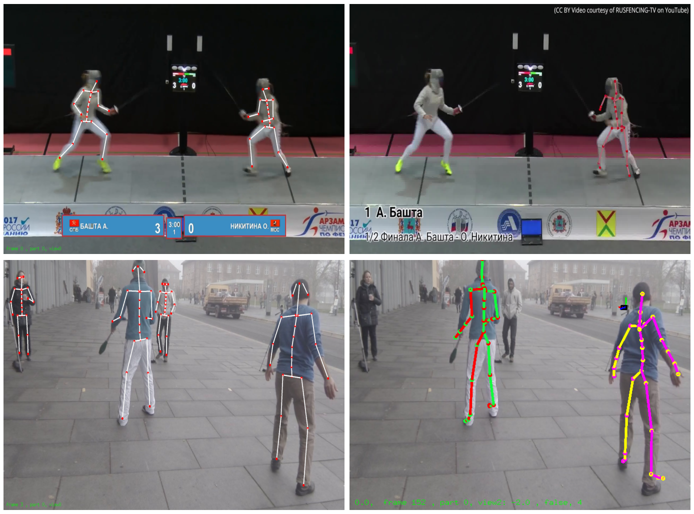


Fig. 7: Side-by-side comparison of our method against the monocular single-person human pose estimation methods of Mehta et al. [38] (top right) and the offline method of Elhayek et al. [18] (bottom right) which tracks two persons with three cameras. Our approach succeeds in accurately tracking all persons in the scene (left column).

ular camera setup as depth estimation is severely ill posed. Thus, a slight inaccuracy in the 2D body-parts estimation leads to big error in the depth estimation. Unlike other methods, our approach is still able to recover from the tracking failures, even after long occlusion of many body-parts; see the supplementary video. Our tracking results of many sequences show that our algorithm succeeds in challenging multi-person scenarios where all other human motion tracking methods based on single RGB camera fail. Moreover, we achieve high temporal stability and reasonable accuracy. This accuracy can also be improved by using 2D body part detector which is more stable to occlusions.

6 Conclusion and Future Work

We have presented the first fully automatic method to estimate 3D kinematic poses of multiple persons in temporally stable manner directly from a single RGB camera. Our approach automatically detects the number of persons in the scene and generates corresponding person-specific 3D skeletons based on anthropometric data tables. It also automatically estimates the initial 3D location of each person which allows to define their coarse initial poses. In the tracking phase, it fits each 3D skeleton to the corresponding 2D body-parts detections. These detections can be estimated using any 2D body-part estimation method which allows to easily upgrade our algorithm with any progress in 2D pose estimation. Our algorithm dynamically generates 3D skeletons for persons who enter the scene and delete the skeletons of those who leave. In contrast to previous works, our fully automatic algorithm can operate with multiple persons in real-time without the need of bounding boxes. This makes our algorithm optimal for VR

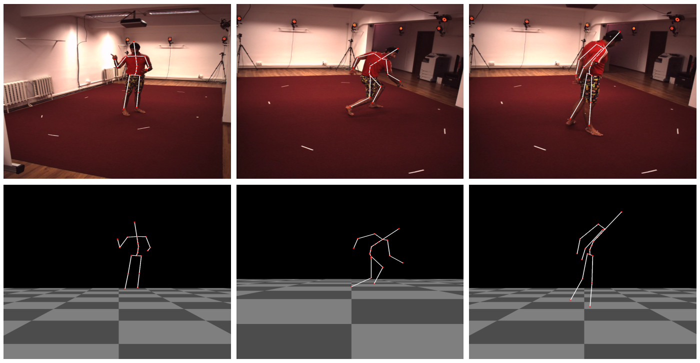


Fig. 8: Sample images from H3.6M sequences used for quantitative evaluations. Top row shows overlaid 2D Skeletons and bottom row shows 3D visualizations of the captured skeletons. From left to right, we show tracking results of *Directions*, *Posing* and *Waiting* sequences for Subject $S9$ whose Mean Per Joint Position Error is $153mm$, $158mm$ and $167mm$ respectively.

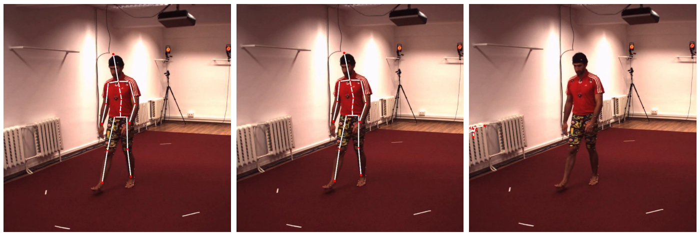


Fig. 9: Importance of algorithmic components. Left: tracking result of our algorithm; MPJPE 90mm. Middle: an alternative of our algorithm constructed by removing the skeleton generation step (i.e. using the default skeleton); MPJPE 460mm. Right: second alternative constructed by removing initial pose localization step which fails completely.

application. We have demonstrated the effectiveness of our system by tracking many sequences with strong distortion in videos, strong illumination changes, and multiple persons performing complex motions. Moreover, we have shown results in real-time scenarios, including live streaming from a webcam. As future work, we are going to investigate the problem of depth estimation uncertainty which could be reduced with domain specific knowledge. Furthermore, in order to improve the run-time of our algorithm, we intend to employ more advanced optimization algorithms.

Acknowledgements

This work has been partially funded by the Federal Ministry of Education and Research of the Federal Republic of Germany as part of the research projects DYNAMICS (Grant number 01IW15003) and VIDETE (Grant number 01IW18002).

References

1. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE transactions on pattern analysis and machine intelligence **28**(1), 44–58 (2006)
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1446–1455 (2015)
3. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: BMVC (2013)
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
5. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proc. ICCV. pp. 1092–1099 (2011)
6. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. CVPR, IEEE (June 2014)
7. Bo, L., Sminchisescu, C.: Twin gaussian processes for structured prediction. IJCV **87**, 28–52 (2010)
8. 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. pp. 561–578 (2016)
9. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: CVPR (2014)
10. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: CVPR. pp. 8–15 (1998)
11. "Bulat, A., Tzimiropoulos, Georgios", e.B., Matas, J., Sebe, N., Welling, M.: Human Pose Estimation via Convolutional Part Heatmap Regression, pp. 717–732. Springer International Publishing, Cham (2016)
12. C. Gordon, C. Blackwell, M.M., Kristensen, S.: 2012 anthropometric survey of u.s. army personnel: Methods and summary statistics (Natick/TR-15/007) (2014)
13. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
14. Charles, J., Pfister, T., Magee, D.R., Hogg, D.C., Zisserman, A.: Personalizing human video pose estimation. CoRR **abs/1511.06676** (2015), http://arxiv.org/abs/1511.06676
15. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B.: Synthesizing training images for boosting human 3d pose estimation. In: 3D Vision (3DV) (2016)
16. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., et al.: Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (TOG) **35**(4), 114 (2016)
17. Du, Y., Wong, Y., Liu, Y., Han, F., Gui, Y., Wang, Z., Kankanhalli, M., Geng, W.: Marker-less 3d human motion capture with monocular image sequence and height-maps. In: European Conference on Computer Vision. pp. 20–36. Springer (2016)
18. Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Marconi: Convnet-based marker-less motion capture in outdoor and indoor scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(3), 501–514 (March 2017). https://doi.org/10.1109/TPAMI.2016.2557779
19. Elhayek, A., 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)

20. Elhayek, A., 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: Proc. CVPR (2015)
21. Elhayek, A., Stoll, C., Hasler, N., Kim, K.I., Seidel, H.P., Theobaltl, C.: Spatio-temporal motion tracking with unsynchronized cameras. In: Proc. CVPR (2012)
22. Elhayek, A., Stoll, C., Hasler, N., Kim, K.I., Theobaltl, C.: Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In: Proc. CGF (2014)
23. Fan, X., Zheng, K., Zhou, Y., Wang, S.: Pose locality constrained representation for 3d human pose reconstruction. In: European Conference on Computer Vision. pp. 174–188. Springer (2014)
24. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture – a multi-layer framework. IJCV **87**, 75–92 (2010)
25. Hasler, N., Rosenhahn, B., Thormählen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: CVPR (2009)
26. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schieke, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision (ECCV) (2016), http://arxiv.org/abs/1605.03170
27. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2014)
28. Jain, A., Thormählen, T., Seidel, H.P., Theobalt, C.: Moviereshape: Tracking and reshaping of humans in videos. ACM Trans. Graph. (Proc. SIGGRAPH Asia 2010) **29**(5) (2010)
29. Kostrikov, I., Gall, J.: Depth sweep regression forests for estimating 3d human pose from images. In: BMVC. vol. 1, p. 5 (2014)
30. Lee, C.S., Elgammal, A.: Coupled visual and kinematic manifold models for tracking. IJCV **87**, 118–139 (2010)
31. Lee, H.J., Chen, Z.: Determination of 3d human body postures from a single view. Computer Vision, Graphics, and Image Processing **30**(2), 148–168 (1985)
32. Leonardos, S., Zhou, X., Daniilidis, K.: Articulated motion estimation from a monocular image sequence using spherical tangent bundles. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on. pp. 587–593. IEEE (2016)
33. Li, R., Tian, T.P., Sclaroff, S., Yang, M.H.: 3d human motion tracking with a coordinated mixture of factor analyzers. IJCV **87**, 170–190 (2010)
34. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision. pp. 332–347. Springer (2014)
35. LI, S., Liu, Z.Q., Chan, A.B.: Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2014)
36. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2848–2856 (2015)
37. Mehta, D., Rhodin, H., Casas, D., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation using transfer learning and improved cnn supervision. arXiv preprint arXiv:1611.09813 (2016)
38. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. In: ACM Transactions on Graphics. vol. 36 (Jul 2017)

39. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU **104**(2), 90–126 (2006)
40. Park, S.W., eun Kim, T., Choi, J.S.: Robust estimation of heights of moving people using a single camera. In: ICITCS (2011)
41. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. CoRR **abs/1611.07828** (2016), http://arxiv.org/abs/1611.07828
42. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4929–4937 (2016)
43. Plankers, R., Fua, P.: Tracking and modeling people in video sequences. CVIU pp. 285–302 (2001)
44. Poppe, R.: Vision-based human motion analysis: An overview. CVIU **108**(1-2), 4–18 (2007)
45. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: Egocentric marker-less motion capture with two fisheye cameras. ACM Trans. Graph. **35**(6), 162:1–162:11 (Nov 2016). https://doi.org/10.1145/2980179.2980235, http://doi.acm.org/10.1145/2980179.2980235
46. Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H.P., Theobalt, C.: General automatic human shape and motion capture using volumetric contour cues. In: European Conference on Computer Vision. pp. 509–526. Springer (2016)
47. Rogge, L., Klose, F., Stengel, M., Eisemann, M., Magnor, M.: Garment replacement in monocular video sequences. ACM Transactions on Graphics **34**(1), 6:1–6:10 (Nov 2014)
48. Shiratori, T., Park, H.S., Sigal, L., Sheikh, Y., Hodgins, J.K.: Motion capture from body-mounted cameras. ACM Trans. Graph. **30**(4), 31:1–31:10 (Jul 2011)
49. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV **87**, 4–27 (2010)
50. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. In: ICCV. pp. 915– 922 (2003)
51. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: ICCV (2011)
52. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3d body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 991–1000 (2016)
53. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660 (2014)
54. Urtasun, R., Fleet, D.J., Fua, P.: Temporal motion models for monocular and multiview 3d human body tracking. Comput. Vis. Image Underst. **104**(2), 157–177 (Nov 2006). https://doi.org/10.1016/j.cviu.2006.08.006, http://dx.doi.org/10.1016/j.cviu.2006.08.006
55. Valmadre, J., Lucey, S.: Deterministic 3d human pose estimation using rigid structure. Computer Vision–ECCV 2010 pp. 467–480 (2010)
56. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)
57. Ye, M., Shen, Y., Du, C., Pan, Z., Yang, R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1517–1532 (Aug 2016). https://doi.org/10.1109/TPAMI.2016.2557783

58. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a CNN coupled with a geometric prior. CoRR **abs/1701.02354** (2017), http://arxiv.org/abs/1701.02354