2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Human Motion Capture Using a Drone

Xiaowei Zhou, Sikang Liu, Georgios Pavlakos, Vijay Kumar, Kostas Daniilidis

***Abstract*— Current motion capture (MoCap) systems generally require markers and multiple calibrated cameras, which can be used only in constrained environments. In this work we introduce a drone-based system for 3D human MoCap. The system only needs an autonomously flying drone with an on-board RGB camera and is usable in various indoor and outdoor environments. A reconstruction algorithm is developed to recover full-body motion from the video recorded by the drone. We argue that, besides the capability of tracking a moving subject, a flying drone also provides fast varying viewpoints, which is beneficial for motion reconstruction. We evaluate the accuracy of the proposed system using our new DroCap dataset and also demonstrate its applicability for MoCap in the wild using a consumer drone.**

I. Introduction

Capturing 3D human body motion is a challenging problem with many applications, e.g., in human-computer interaction, health care and sports. This problem has been conditionally solved by multi-camera motion capture (MoCap) systems (e.g. Vicon and Qualysis) in constrained studios. However, those MoCap systems suffer from their inflexibility and inconvenience: cameras require reliable fixation and frequent calibration, the tracking space is limited and fixed, and the subject should wear special markers. While being more challenging, image-based MoCap with an RGB camera has wider applicability and draws an increasing attention in recent years.

Despite the remarkable advances in monocular 3D human pose estimation (see the related-work section), these methods suffer from the inherent ambiguity of single-view reconstruction. The ambiguity is alleviated by learning a 3D pose prior from existing MoCap datasets but cannot be resolved geometrically. Another line of work aims to leverage multi-frame information in a video to reconstruct a nonrigid shape, which is known as nonrigid structure from motion (NRSFM) [1]. However, NRSFM requires sufficiently fast camera motion relative to the object [2], [3], which is impractical if the camera is fixed.

To address the above limitations of previous approaches, we propose a novel system for human body MoCap using a drone (see Figure 1) leveraging the state-of-the-art techniques in autonomous drones and computer vision. An autonomously flying drone orbits and records a video of the subject, providing fast varying viewpoints about the subject. A convolutional neural network (CNN) based 2D pose estimator produces reliable 2D tracks of body joints from the video, which are input to a 3D pose estimator that robustly initializes reconstruction and suppresses outliers in the 2D tracks. Finally, a NRSFM algorithm is developed to further refine the reconstruction using sequence information and impose the articulation constraint of human body.

Fig. 1. We propose a novel system for human motion capture based on an autonomously flying drone. (left) The drone orbits the human subject and records a video with an on-board RGB camera. The 3D full-body pose is recovered from the monocular video. (right) The reconstructed pose in the example frame is visualized at a novel viewpoint. The original viewpoint from the drone is indicated by the square pyramid. Please see the video at https://github.com/daniilidis-group/drocap.

Our contributions are summarized below:

- We propose a novel drone-based system for human MoCap, which is simple (uses only a drone with an on-board RGB camera), flexible (works both indoors and outdoors) and readily usable (needs no particular system calibration or model training).
- We argue that, compared to using a static camera, using a drone for video recording is able to provide fast camera motion relative to the subject, which is necessary for motion reconstruction.
- We develop a reconstruction algorithm to recover 3D human poses from the drone-based video, which consists of a novel synthesis between single-view pose estimation and sequence-based NRSFM for both robustness and accuracy.
- We introduce a novel rank-1 constraint to model the articulation of human body. The proposed constraint is effective and applicable even if the limb lengths of the subject are unknown.
- We release a drone-based MoCap dataset named DroCap, which consists of human motion videos captured by a flying drone and ground truth obtained by a Vicon MoCap System. The dataset is available at https://github.com/daniilidis-group/drocap.

X. Zhou is with the the State Key Laboratory of CAD&CG, Zhejiang University. Email address: xzhou@cad.zju.edu.cn.

S. Liu, G. Pavlakos, V. Kumar and K. Daniilidis are with the GRASP laboratory, University of Pennsylvania. Email addresses: {sikang,pavlakos,kumar,kostas}@seas.upenn.edu.

Related work

Markerless human motion capture has been a long standing problem in computer vision. Early work in this area was focused on model-based body tracking in monocular [4] or multi-view sequences [5], which in general requires initialization in the first frame and is prone to tracking failures. To address these issues, more robust bottom-up approaches were proposed. These approaches first detect 2D body joints in images and lift them to 3D by assuming a given skeleton [3], [6], [7], searching MoCap databases [8] or learning statistical models from MoCap data, such as the loosely-limbed model [9], principal component analysis [10]–[12], sparse representation [13]–[15] and body shape models [16], [17]. The main limitation of these approaches is that single-view reconstruction is inherently ill-posed and using geometric priors alone is insufficient to resolve the reconstruction ambiguity. Moreover, the pose prior learned from training data might not be able to explain the pose variability in testing examples resulting in inaccurate reconstructions. While the reconstruction ambiguity can be resolved by using multiple cameras [18]–[20], the calibrated multi-camera system is not easily accessible in practice.

Another strand of work directly learns the mapping from an image to 3D pose parameters using annotated training data in the form of image-pose pairs, which is referred to as discriminative methods. The advantage of discriminative methods is their ability to leverage image cues, such as shading and occlusion, to reduce the reconstruction ambiguity. A variety of supervised learning methods have been adopted ranging from traditional techniques such as linear regression [21], kernel density estimation [22] and Gaussian processes [23] to modern deep convolutional neural networks (CNNs) [24], [25]. The main limitation for the discriminative methods is that they require a large number of training images with corresponding 3D pose annotations, which could only be collected using MoCap systems. While there have been large-scale MoCap datasets, such as HumanEva [26] and Human3.6M [22], they lack diversity in scenes, subject appearances and actions. As a consequence, the models trained on indoor MoCap datasets are prone to overfitting and can hardly be applied to outdoor images. Some recent works tried to address the scarcity of training data by image synthesis [27], [28], but the performance of a model trained on synthesized images is not guaranteed when applied to real images due the difference in image statistics.

Human motion capture is closely related to nonrigid structure from motion (NRSFM) [1], [29]–[32]. In NRSFM, a deformable shape is reconstructed from 2D keypoint tracks extracted from a video. Most of the state-of-the-art NRSFM methods assume a low-rank prior on the deformable shape over time. But unlike the single-view methods that learn bases from MoCap data (e.g. [10], [13]), NRSFM methods recover the bases during reconstruction without the need of training data, which might better fit the subject. However, it is difficult to apply existing NRSFM pipelines for human motion capture. First, obtaining clean 2D keypoint tracks from a video is difficult especially for fast moving and deformable objects like human body. Second, NRSFM requires fast camera motion [2], [3], which is impractical in usual scenarios. In this work, we leverage NRSFM but combine it with single-view pose estimation for robust initialization and outlier handling. Moreover, using an on-board camera on a drone for video recording naturally provides fast camera motion.

Using flying cameras for human MoCap was proposed in [33], in which the system consists of multiple drones equipped with RGB-D sensors and solves the problem by model-based tracking. However, it requires a scanning stage in which the subject stays static for body shape scanning using depth sensors. Also, RGB-D sensors are restricted to indoor environments and the sensing range is limited. Compared to [33], the proposed system is more widely usable. It only needs a consumer drone with an RGB camera, doesn't require any initialization or scanning stage, and works in both indoor and outdoor environments.

II. Approaches

An autonomously flying drone is used for data collection. The drone tracks and orbits the subject with an on-board RGB camera pointing at the subject and recording a video. This functionality has been implemented in many commercial drones such as DJI Phantom 4 and Mavic Pro. The motivation for using a drone instead of a fixed camera for video recording is the capability to provide a sequence of fast varying viewpoints about the subject, which is favorable to motion reconstruction. The importance of camera motion will be experimentally demonstrated in Section III-A.

Given a monocular video of the subject recorded from the orbiting drone, we aim to recover the 3D pose sequence of the subject. Our pipeline consists of the following steps: 1) *2D pose detection* in which the subject is detected and the 2D pose is estimated in each frame; 2) *single-frame initialization* in which the camera viewpoints and 3D human poses are initialized by the single-view pose estimation method [34]; 3) *multi-frame bundle adjustment* in which the camera viewpoints and 3D poses are refined by minimizing the nuclear norm of shape matrix with an articulation constraint. This pipeline is analogous to the successful experience in rigid structure from motion: we detect keypoints, incrementally reconstruct each frame and optimize all unknowns in bundle adjustment.

A. 2D pose detection

The bounding box of the subject in each frame is obtained by either object tracking or detection. For example, the DJI Mavic Pro comes with the active tracking feature and provides the bounding box of the tracked subject. Otherwise, an object detector can be applied to localize the subject in each frame, e.g., the faster R-CNN [35] in our experiments.

We adopt the stacked hourglass model proposed by Newell et al. [36] for 2D pose estimation, which is the state-of-the-art method for this problem. It is a fully convolutional neural network (F-CNN) with two stacked hourglass components,

each of which consists of a series of downsampling and upsampling layers implementing the bottom-up and top-down processing to integrate local image feature with global context over the whole image. The input to the network is a 2D image with a bounding box around the subject and the output is a set of heat maps with each showing the predicted spatial distribution of the corresponding joint. The details are referred in [36].

B. Single-frame initialization

After the 2D body joints are located in the image sequence, NRSFM methods can be used to reconstruct the 3D poses from the 2D tracks of body joints. However, there are two difficulties for this approach. First, there might be a considerable number of gross errors (outliers) in the detected 2D joints due to occlusion, left-right ambiguity and background clutters. The existing NRSFM methods can hardly handle outliers as NRSFM is an ill-posed problem without much information to correct the input error. Secondly, NRSFM methods typically rely on low-rank factorization which requires a predefined rank while the best rank is often unknown.

To address these difficulties, we propose to use a recent method for single-view 3D pose estimation [34] to initialize the reconstruction. In [34], a pose dictionary is learned from existing MoCap data and the pose to be reconstructed is assumed as a linear combination of the bases in the dictionary. In this way, the number of unknowns is reduced and the optimization is easier to solve compared to NRSFM where the bases are also unknown. An EM algorithm is also developed in [34] to account for uncertainties in CNN based 2D joint localization by incorporating the 3D pose prior learned from MoCap data. Even if the learned pose dictionary cannot perfectly represent the poses to be reconstructed, this step can reliably initialize the camera viewpoints and correct outliers in the 2D input.

C. Multi-frame bundle adjustment

A downside of the single-view reconstruction is that the pose bases learned from other MoCap datasets might not be able to represent the new poses in test images. We propose to solve this problem by adapting the pose bases to the current sequence, which has been realized in NRSFM where a low-rank basis is learned along with other unknowns from 2D keypoint tracks.

We adopt the nuclear norm minimization scheme which has been used in many recent NRSFM methods (e.g. [30], [31]). Compared to factorization based methods, the advantages are two-fold: 1) there is no need to explicitly define a rank; and 2) nuclear norm minimization is convex. We additionally propose a novel rank-1 constraint to model the articulated nature of human body.

1) Objective function: Suppose the 2D pose and 3D pose of the subject in frame t are represented by $W_t \in \mathbb{R}^{2\times p}$ and $S_t \in \mathbb{R}^{3\times p}$ respectively, where p is the number of joints. Following the general practice in NRSFM (e.g. [1], [30]), we assume that an orthographic camera model is used and both W_t and S_t are centralized, such that

$$W_t = R_t S_t \tag{1}$$

where $R_t \in \mathbb{R}^{2\times 3}$ denotes the first two rows of the camera rotation matrix at frame t.

Given the 2D pose sequence $W = \{W_1, \cdots, W_n\}$, we recover the 3D pose sequence $S = \{S_1, \cdots, S_n\}$ and the camera rotation sequence $R = \{R_1, \cdots, R_n\}$ by solving the following optimization problem:

$$\min_{S,R,L} f(S,R,L) + \alpha \|S^{\#}\|_* \tag{2}$$

where $f(S,R,L)$ is a smooth function composed of the following terms:

$$f(S,R,L) = \sum_{t=1}^{n} \|W_t - R_t S_t\|_F^2 + \gamma \|\ell(S) - L\|_F^2 \tag{3}$$

The first term in (3) is the sum of reprojection errors over all joints in all frames. The second term enforces the articulation (anthropomorphic) constraint, i.e., the limb lengths should be constant across the sequence. However, as the scale of the reconstructed 3D structure is determined by the given 2D structure under the orthographic projection, the size of the reconstructed structure may vary in different frames depending on the distance from the camera to the subject. Therefore, instead of constraining limb lengths as constants, we enforce that the the ratios between limb lengths to be unchanged across frames. Suppose $\ell(\cdot)$ is a function such that the t-th column of $\ell(S)$ gives the squared limb lengths of S_t, $\ell(S)$ should be rank-1 if the articulation constraint is satisfied. To simplify the optimization, we minimize the difference between $\ell(S)$ and an auxiliary rank-1 matrix L instead of directly constraining the rank of $\ell(S)$. L is also unknown and updated during optimization. Note that $\ell(S)$ gives the squared lengths which are differentiable.

The second term in (2) is a nonsmooth regularizer that enforces the low-rankness of the reconstructed poses, where $\|\cdot\|_*$ is the nuclear norm and $S^{\#}$ denotes a rearrangement of S such that the t-th column of $S^{\#}$ is the vectorized S_t. When $\gamma = 0$, the formulation (2) is identical to the ones used in previous work for NRSFM (e.g. [30], [31]).

Note that the temporal smoothness of both S_t and R_t could be conveniently imposed by minimizing their temporal changes. We didn't include them in (2) for simplicity and observed that adding them didn't significantly improve the quantitative results.

2) Optimization: The problem in (2) is nonlinear and nonconvex. However, the single-frame initialization stage has provided reliable initialization, which allows us to solve the problem in (2) by local optimization.

More specifically, we alternately update each variable while fixing the others. The camera rotation R can be updated with any parameterization of rotation matrix. In accordance with the initialization method [34], we optimize R over the Stiefel manifold, which is implemented using

the manifold optimization toolbox [37]. The update of L is a standard low-rank approximation problem which can be analytically solved by singular value decomposition. The update of S is more complicated where the objective consists of a smooth loss function and a nonsmooth nuclear norm regularizer. We adopt the proximal gradient (PG) method [38], which updates S iteratively according to the following rule until convergence:

$$S^{k+1} = \arg\min_{S} \frac{1}{2} \left\| S - [S^k - \frac{1}{\mu} \nabla f_{S^k}] \right\|_F^2 + \frac{\alpha}{\mu} \left\| S^{\#} \right\|_* \tag{4}$$

where ∇f_{S^k} is the gradient of the smooth function f evaluated at the previous estimate S^k and μ determines the step size. The subproblem in (4) can be analytically solved by the singular value thresholding [39]. To additionally accelerate the convergence of the PG iterations, the Nesterov acceleration scheme [38] is also implemented.

III. Results

A. Importance of camera motion

We first demonstrate that fast camera motion is favorable to motion reconstruction, which is the motivation of using drone for data collection in the proposed system. To achieve an arbitrary camera velocity, we use synthetic 2D input by projecting groundtruth 3D body joints to 2D with a virtual orthographic camera rotating around the subject. The 3D data is from Human3.6M [22], a large-scale MoCap dataset. All sequences of 15 actions from Subject 9 and 11 are used for evaluation. The sequences are subsampled at a frame rate of 24 fps and the first 10 seconds of each sequence are used for evaluation. The reconstruction error is used as the error metric:

$$e = \min_{\mathcal{T}} \frac{1}{p} \sum_{i=1}^{p} \|\hat{\boldsymbol{x}}_i - \mathcal{T}(\boldsymbol{x}_i^*)\|_2,$$

which calculates the mean distance between the estimated joint locations $\hat{\boldsymbol{x}}$ and ground truth $\boldsymbol{x}^*$ after a similarity transformation $\mathcal{T}$ to align them.

The mean reconstruction error averaged over all sequences is plotted in Figure 2 as a function of the angular velocity of the virtual camera. The results of the initial single-view method [34] and the bundle adjustment with ($\gamma = 1$) and without ($\gamma = 0$) the articulation constraint are presented. To avoid training on the same dataset, we learn a pose dictionary from the CMU Mocap dataset [40] for single-frame initialization, achieving a mean error around 77mm. If the pose dictionary is learned from the same dataset (the training set of Human3.6M), the mean error is around 48mm, but this setting is impractical in real applications. Figure 2 shows that the multi-frame bundle adjustment improves upon the initial single-view estimation and the accuracy becomes better as the camera rotates faster, which validates that the multi-view information from fast varying viewpoints helps motion reconstruction. The benefit of imposing the articulation constraint is also clearly demonstrated.

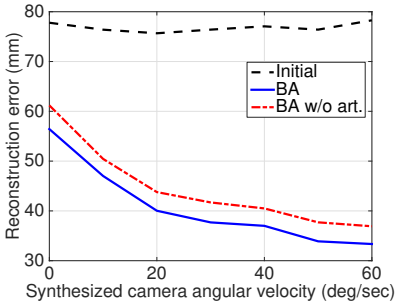

Fig. 2. The 3D reconstruction error as a function of the angular velocity of virtual camera. The three curves correspond to the single-view initialization (Initial) by [34], the multi-frame bundle adjustment (BA) and BA without the articulation constraint (BA w/o art.).

B. DroCap dataset

We collect a new drone-based MoCap dataset named DroCap. The ground truth of body motion is provided by a Vicon MoCap system with markers attached to body joints. We mount a GoPro camera on the AscTec Pelican quadrotor platform (Figure 3) and program it to autonomously track a pre-defined trajectory which is a circle centered at the subject and the desired orientation of the camera is always pointing at the center. The desired speed is 1m/s corresponding to an angular velocity around $25^\circ/s$. During data collection, the subject is doing a variety of actions, such as walking, boxing, and playing soccer, staying at the same location due to the limited indoor space. The current dataset consists of 6 sequences from 2 subjects.

The 3D human poses reconstructed from the monocular videos are compared to the ground truth obtained from Vicon. Note that no training data is provided. For the proposed approach, the stacked hourglass model [36] trained on MPII [42] is adopted for 2D pose estimation and a pose dictionary learned from Human3.6M [22] is used for single-frame initialization.

The qualitative results on several representative frames are visualized in Figure 4. While the initial single-view estimates by [34] have captured global structures, the reconstructions after the multi-frame bundle adjustment are closer to the ground truth recovering more faithful details, e.g., the joint angles of elbows or knees. The bottom-left figure in Figure 4 shows an example where the original 2D pose estimate is inaccurate but the final reconstruction is correct after handling 2D uncertainties by [34].

The reconstruction errors at 12 joints (wrists, elbows, shoulders, hips, knees and ankles) are evaluated. The mean reconstruction errors for each sequence are given in Table I. "Initial" and "BA" denote single-frame initialization and multi-frame bundle adjustment, respectively. A baseline method "MF + NNM" is included in comparison, where the 2D joint tracks detected by the same CNN-based detector are input to the state-of-the-art NRSFM method, i.e., matrix factorization for initialization followed by nuclear norm minimization for structure refinement [30]. The proposed approach outperforms the baselines, achieving an average

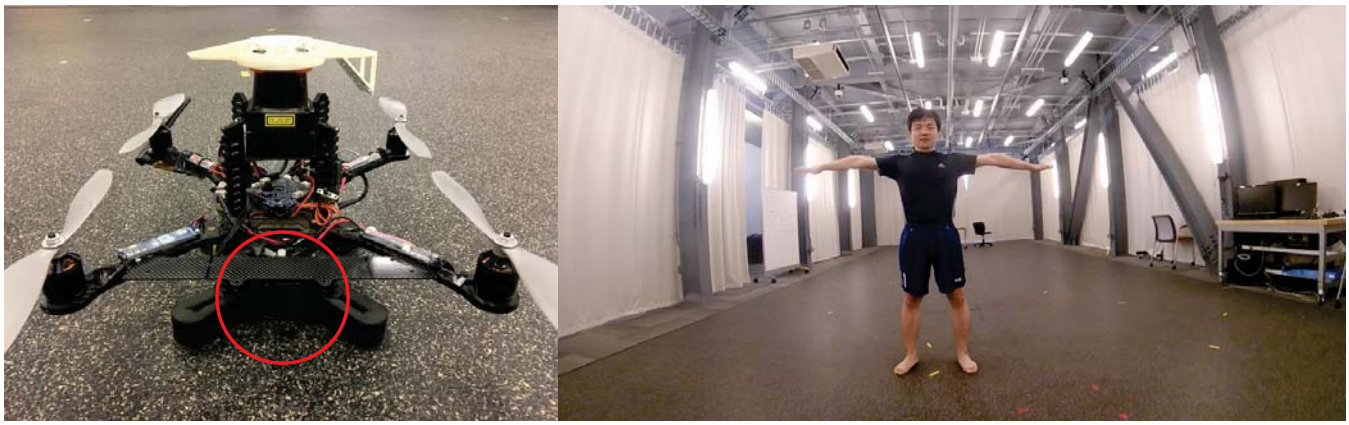

Fig. 3. Left: the AscTec Pelican quadrotor platform used for data collection (the red circle labels the on-board GoPro camera). Right: a sample video frame from the on-board camera.

Fig. 4. Qualitative evaluation on the DroCap dataset. Each panel corresponds to an example frame. For each panel, the 1st column shows the cropped image (top) and estimated 2D pose (bottom). The 2nd to the 4th columns correspond to the groundtruth poses (GT), reconstructions after bundle adjustment (BA) and from the initialization (Initial), respectively, visualized in front and profile views.

Fig. 5. Reconstruction from outdoor sequences captured by DJI Mavic Pro. An original frame of each sequence is shown in the top row. The following rows correspond to several selected frames, showing the estimated 2D pose in the cropped image and the 3D pose visualized from the front and profile views. The last sequence is from a YouTube video [41].

TABLE I

THE MEAN RECONSTRUCTION ERRORS (MM) ON THE DROCAP DATASET.

	Box1	Box2	Walk1	Walk2	Soccer1	Soccer2	Mean
MF+NNM [30]	57.3	86.4	78.1	63.2	123.9	93.2	83.7
Initial [34]	74.0	86.7	62.0	77.0	75.7	78.4	75.6
Initial+BA	**53.9**	**70.6**	**41.1**	**47.2**	**56.3**	**62.6**	**55.3**

TABLE II

THE MEAN RECONSTRUCTION ERRORS (MM) FOR DIFFERENT JOINTS.

Wrist	Elbow	Shoulder	Hip	Knee	Ankle
70.4	62.8	39.1	39.5	57.6	62.2

error around 55mm. The performance gain over the single-frame initialization is mainly due to the existence of fast camera motion in drone-based videos that provides richer information for reconstruction. Moreover, the gain is more significant for more repetitive motions such as walking as the pose sequence can be better represented by a low-dimensional subspace. The errors for separate joints are presented in Table II.

C. Outdoor MoCap

Finally, we demonstrate the applicability of the proposed system for outdoor MoCap using consumer drone DJI Mavic Pro. The built-in active tracking function on Mavic Pro is used to track and orbit the moving subject autonomously. Several example sequences are shown in Figure 5, including a YouTube video [41] to demonstrate the generalizability of the proposed algorithm. The same as previous experiments, no additional training is used. The generic stacked hourglass model [36] trained on MPII [42] is used for 2D pose estimation and the pose dictionary learned from Human3.6M [22] is used for single-frame initialization. The reconstruction results for several selected frames are shown in Figure 5. As shown, the details of the subject motion are well captured.

For example, we can clearly see in the last sequence that the right arm of the subject is swinging while the left hand that holds the remote controller is relatively static.

D. Running time

The reconstruction algorithm was running offline on a desktop with an Intel i7 3.4G CPU, 8G RAM and a GeForce GTX Titan X 6GB GPU. The running time per frame was $\sim$0.2s for 2D pose estimation and $\sim$0.3s for single-frame initialization, which could be easily paralleled. For a sequence of 300 frames, the running time for multi-frame bundle adjustment was $\sim$8s.

IV. Discussion

We proposed a novel system for human MoCap using an autonomously flying drone, aimed to address limitations of existing MoCap systems that rely on markers and static cameras. The proposed system is applicably both indoors and outdoors and is capable of using a consumer drone without the need of particular system calibration or model training. We also introduced a new dataset for drone-based MoCap. This work is an initial effort towards drone-based MoCap, which can be potentially extended, e.g. using multiple drones or active trajectory planning, for more accurate reconstruction.

References

[1] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *CVPR*, 2000.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *PAMI*, vol. 33, no. 7, pp. 1442–1456, 2011.

[3] H. S. Park and Y. Sheikh, "3D reconstruction of a smooth articulated trajectory from a monocular image sequence," in *ICCV*, 2011, pp. 201–208.

[4] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *CVPR*, 2003.

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

[6] C. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," *CVIU*, vol. 80, no. 3, pp. 349–363, 2000.

[7] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *PAMI*, vol. 28, no. 7, pp. 1052–1062, 2006.

[8] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *CVPR*, 2016.

[9] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation," *IJCV*, vol. 98, no. 1, pp. 15–48, 2012.

[10] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single Image 3D Human Pose Estimation from Noisy Observations," in *CVPR*, 2012.

[11] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A Joint Model for 2D and 3D Pose Estimation from a Single Image," in *CVPR*, 2013.

[12] F. Zhou and F. D. la Torre, "Spatio-temporal matching for human detection in video," in *ECCV*, 2014.

[13] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *ECCV*, 2012.

[14] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *CVPR*, 2015.

[15] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3d shape estimation: A convex relaxation approach," *PAMI*, vol. 39, no. 8, pp. 1648–1661, 2017.

[16] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *ICCV*, 2009.

[17] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *ECCV*, 2016.

[18] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in *CVPR*, 2014.

[19] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys, "Joint camera pose estimation and 3D human pose estimation in a multi-camera setup," in *ACCV*, 2014.

[20] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras," in *CVPR*, 2015.

[21] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *PAMI*, vol. 28, no. 1, pp. 44–58, 2006.

[22] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *PAMI*, vol. 36, no. 7, pp. 1325–1339, 2014.

[23] L. Bo and C. Sminchisescu, "Twin Gaussian processes for structured prediction," *IJCV*, vol. 87, no. 1-2, pp. 28–52, 2010.

[24] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3D human pose estimation," in *ICCV*, 2015.

[25] B. Tekin, A. Rozantsev, , V. Lepetit, and P. Fua, "Direct prediction of 3D body poses from motion compensated sequences," in *CVPR*, 2016.

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

[27] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen *et al.*, "Synthesizing training images for boosting human 3D pose estimation," in *3DV*, 2016.

[28] G. Rogez and C. Schmid, "MoCap-guided data augmentation for 3D pose estimation in the wild," in *NIPS*, 2016.

[29] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito, "Factorization for non-rigid and articulated structure using metric projections," in *CVPR*, 2009.

[30] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *IJCV*, vol. 107, no. 2, pp. 101–122, 2014.

[31] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *CVPR*, 2013.

[32] B. Wandt, H. Ackermann, and B. Rosenhahn, "3d reconstruction of human motion from monocular image sequences," *PAMI*, vol. 38, no. 8, pp. 1505–1516, 2016.

[33] L. Xu, L. Fang, W. Cheng, K. Guo, G. Zhou, Q. Dai, and Y. Liu, "Flycap: Markerless motion capture using multiple autonomous flying cameras," *arXiv preprint arXiv:1610.09534*, 2016.

[34] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *CVPR*, 2016.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." *PAMI*, 2016.

[36] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[37] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *JMLR*, vol. 15, pp. 1455–1459, 2014.

[38] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Tech. Rep., 2007.

[39] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[40] "Mocap: Carnegie mellon university motion capture database. http://mocap.cs.cmu.edu/."

[41] "DJI Mavic Pro active track – trace and profile feature. https://www.youtube.com/watch?v=XiAL8hMccdc&t."

[42] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.