Human Motion Capture Using a Drone

Zhang, Liqiang

July 10, 2018

Abstract

Current motion capture (MoCap) system generally require markers and multiple calibrated cameras, which can be used only in constrained environments. In this paper, the author 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. The author think that besides the capability of tracking a moving subject, a flying drone. Besides the capability of tracking a moving subject, a flying drone also provides fast varying viewpoints, which is beneficial for motion reconstruction.

1. 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 Qujalysis) in constrained studios [3]. 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.

Depite 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 [1]. 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). However, NRSFM requires sufficiently fast camera motion relative to the ovject, which s impractical if the camera is fixed.

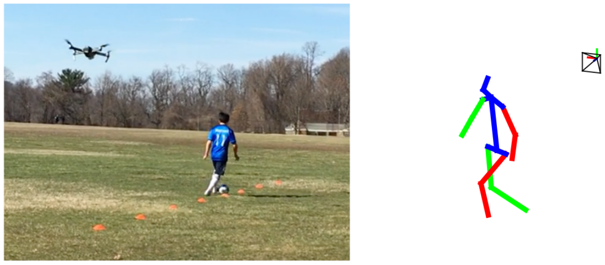

Figure 1. The drone orbits the human subject and records a video with an on-board RGB camera (left). The 3D full-body pose is recovered from the monocular video. The reconstructed pose in the example frame is visualized at a novel viewpoint (right)

To address the above limitations of previous approaches, we propose a novel system for human body MoCap using a drone (Fig. 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 2 tracks of body joints from the video, which are imput 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 [2].

2. Approaches

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. An orthographic camera model is used and both W_t and S_t are centralized in Eq. 1

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

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

	Box1	Box2	Walk1	Walk2	Soccer1	Soccer2	Mean
MF+NNM [4]	57.3	86.4	78.1	63.2	123.9	93.2	83.7
Initial	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 1. The mean reconstruction errors (MM) on the DroCap dataset

$W = \{W_1, \cdots, W_n\}$, this team 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 in Eq. 2

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

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

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

The first term in Eq. 3 is the sun of reprojection errors over all joints in all frames. The second term enforces the articulation (anthropomorphic) constraint, i.e., the limb lenths should be 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 ratios between limb lengths to be unchanged across frames. Suppose $\ell(\text{cot})$ is a function such that the t-th column of $\ell(S)$ gives the squared limb lenths 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.

3. DroCap dataset

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 1. "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.

References

[1] P. Guan, A. Weiss, A. O. Blan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, pages 1381–1388, 2010. 1

[2] M. Paladini, A. D. Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric projections. In *CVPR*, pages 2898–2905, 2009. 1

[3] E. Simo-Serra, A. Ramisa, G. Aleny, and C. Torras. Single image 3d human pose estimation from noisy observations. In *CVPR*, pages 2673–2680, 2012. 1

[4] B. Wandt, H. Ackermann, and B. Rosenhahn. 3D reconstruction of human motion from monocular image sequences. *IEEE TPAMI*, 38(8):1505, 2016. 2