3D Human Reconstruction in the Wild with Collaborative Aerial Cameras

Cherie Ho[1], Andrew Jong[1], Harry Freeman[1], Rohan Rao[1], Rogerio Bonatti[1], Sebastian Scherer[1]

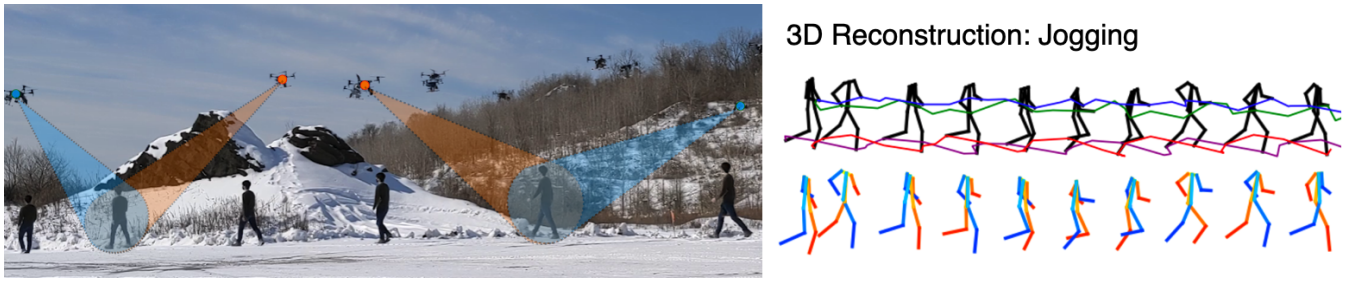


Fig. 1: We present a multi-UAV system for 3D human reconstruction in the wild. Our framework coordinates the motion of multiple aerial cameras to optimally reconstruct the dynamic target's 3D body pose while avoiding obstacles and occlusions. We deploy the system in challenging real-world conditions and target motions such as jogging and playing soccer.

***Abstract*—**

Aerial vehicles are revolutionizing applications that require capturing the 3D structure of dynamic targets in the wild, such as sports, medicine and entertainment. The core challenges in developing a motion-capture system that operates in outdoors environments are: (1) 3D inference requires multiple simultaneous viewpoints of the target, (2) occlusion caused by obstacles is frequent when tracking moving targets, and (3) the camera and vehicle state estimation is noisy. We present a real-time aerial system for multi-camera control that can reconstruct human motions in natural environments without the use of special-purpose markers. We develop a multi-robot coordination scheme that maintains the optimal flight formation for target reconstruction quality amongst obstacles. We provide studies evaluating system performance in simulation, and validate real-world performance using two drones while a target performs activities such as jogging and playing soccer.
Supplementary video: https://youtu.be/jxt91vx0cns

I. Introduction

3D reconstruction of scenes with stereo cameras, RGB-D sensors and monocular cameras is a topic intensively studied in the computer vision and robotics communities [1]–[3]. However, most works focus on static environments, and can only reconstruct static objects. Until recently, capturing dynamic scenes could only be achieved using body markers and high-precision motion capture systems [4], [5], or by markerless systems that heavily rely on skeleton models [6]. Pan-optic studios, on the other hand, rely on visual data from a large number of static cameras to precisely capture motions of multiple targets [7]–[11]. However, they require expensive structures and are confined to small indoor areas.

Aerial camera technologies can significantly extend the capabilities of recording setups to handle dynamic targets in natural outdoor environments. Several works allow drones to detect, track and follow targets in real-time using single-camera systems [12]–[19]. We also find a rich history of work on multi-camera aerial systems that allow users to estimate poses of moving targets. For instance, [20] uses multiple drones to capture the pose of a human wearing markers. [21] explicitly optimizes for body pose reconstruction, but in obstacle-free environments. Most related to our work, [22] presents an aerial motion capture system that uses multi-robot formation controller [23] for data collection. The multi-robot controller avoids obstacles while maintaining formation, but it plans for a short time horizon (1.5s).

Despite the recent progress in multi-drone recording systems, existing approaches are still not able to simultaneously handle all major challenges related to 3D dynamic pose reconstruction in natural environments without markers. A robust system for 3D inference must be able to coordinate the simultaneous recording of various viewpoints of the target even within the presence of obstacles, which can cause image occlusions and robot collisions. In addition, the robots need to navigate smoothly to avoid state estimation noise. Finally, when filming a dynamic target in complex environment, the robot needs to anticipate future actor motions and compute long horizon plans for more optimal motions.

As seen in Figure 1, we tackle these challenges by building upon our previous work in multi-drone cinematography [24], to which a new formation control strategy is introduced for the human reconstruction problem. The proposed system can plan over long horizons multi-robot trajectories for human reconstruction while avoiding occlusions and obstacles. Our contributions are three-fold:

1) Multi-camera coordination: We formulate a multi-camera coordination scheme with the goal of maximizing the reconstructed 3D pose quality of dynamic targets. We develop a scalable two-stage system with long planning time horizons and real-time performance that uses a centralized

[1]The Robotics Institute, Carnegie Mellon University, Pittsburgh PA {cherieh, ajong, hfreeman, rgrao, rbonatti, basti}@cs.cmu.edu

planner for formation control and a decentralized trajectory optimizer that runs on each robot;

2) System scaling and error ablations: We provide extensive simulation experiments validating system performance under different operating conditions such as scaling over multiple drones, and reconstruction error analysis for different magnitudes of camera pose uncertainty.

3) Real-world experiments: We deploy the system in real-world settings using two robots while tracking an actor performing activities such as jogging and playing soccer. We empirically show the improvement in reconstruction quality caused by our adaptive formation scheme.

II. Related Work

Aerial Cinematography and Active Tracking: There is a significant body of work in academia and in industry within the domain of single-drone cinematography. For instance, [12], [13], [15] compute smooth aerial camera plans given user-defined artistic guidelines. As well, commercial drones from Skydio [25] and DJI [26] can track and film actors in complex cluttered environments. Recently, there is growing interest in coordinating multiple UAVs to add viewpoint diversity, with pioneering work that proposes online path planning with inter-drone collision avoidance for indoor settings [27]. [28]–[32] focus on coordinating multi-drone, human-guided shot execution under various constraints, such as smoothness, battery life, and mutual camera visibility. Our previous work [24] increases practicality for multi-UAV tracking of unscripted targets by removing the need for predefined shots while maximizing 3D shot diversity online.

Aerial Motion Capture: Traditionally, human motion capture is achieved by tagging targets with body markers, where recent progress lies in accurate markerless reconstruction using a large number of static visual cameras [9]. Subsequent works investigate mobile UAVs as an alternative to overcome the complexity of this static setup, with [33]–[35] focusing on optimal camera plans for a single UAV. We find works that plan optimal viewpoints with multiple vehicles [21], [36]; however, they do not consider obstacle avoidance. Most related to our work, [23] introduces a decentralized multi-UAV coordination framework for actor position estimation that is extended as a data collection system for outdoor human shape estimation [22]. The system plans for a short horizon (1.5s) to avoid obstacles while maintaining formation around target. However, in our previous work [12], we have shown that planning with longer horizons minimizes the likelihood of myopic trajectories for better target tracking in complex environments. In this work, we present a multi-UAV motion capture system that plans smooth trajectories over a long time horizon (10s) to maintain optimal flight formation for target reconstruction among obstacles.

III. Problem Definition

Our overall goal is to control a team of aerial cameras to reconstruct the 3D pose of a dynamic human moving through a cluttered environment. Let $\theta(t) \in \mathbb{R}^{P\times 3}$ be a vector containing the target's 3D coordinates for P joints at time t. Our mathematical objective is to minimize the reconstruction error E_{recon} calculated with respect to the true target joints over time:

$$E_{\text{recon}} = \sum_{t=1}^{T} ||\hat{\theta}(t) - \theta(t)||^2 \tag{1}$$

In order to minimize this objective (Eq. 1), we capture the scene using a set of aerial cameras, and calculate their trajectories using an optimization framework. Similarly to [12], we employ a weighted set of cost functions that balance robot safety and motion smoothness against visual occlusions of the actor from obstacles. In addition, we introduce a new objective to encode multi-camera collaboration that strives to keep an optimal drone formation over time.

Let $\xi_{qi} : [0, t_f] \rightarrow \mathbb{R}^3 \times SO(2)$ be the trajectory of the i-th UAV, i.e., $\xi_{qi}(t) = \{x(t), y(t), z(t), \psi_q(t)\}$, and $\Xi = \{\xi_{q1}, ..., \xi_{qn}\}$ be the set of trajectories from n UAVs. Let $\xi_a : [0, t_f] \rightarrow \mathbb{R}^3$ be the trajectory of the actor, i.e., $\xi_a(t) = \{x(t), y(t), z(t)\}$, which is inferred using onboard cameras. Let grid $\mathcal{G} : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a voxel occupancy grid that maps every point in space to a probability of occupancy. Let $\mathcal{M}(\mathcal{G}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ be the signed distance values of a point to the nearest obstacle. Each objective is represented as follows:

1) Smoothness: Penalizes jerky motions that may lead to camera blur and unstable flight. Calculated as the sum of costs from individual trajectories: $J_{\text{smooth}}(\Xi) = \sum_i J_{\text{smooth}}(\xi_{qi})$;
2) Occlusion: Penalizes occlusion of the actor by obstacles in the environment for each camera: $J_{\text{occlusion}}(\Xi) = \sum_i J_{\text{occlusion}}(\xi_{qi}, \xi_a, \mathcal{M})$;
3) Obstacle: Penalizes proximity to obstacles that are unsafe for each UAV: $J_{\text{obstacle}}(\Xi) = \sum_i J_{\text{obstacle}}(\xi_{qi}, \mathcal{M})$;
4) Formation: Ensures that the camera formation remains at the optimal configuration for actor reconstruction. Calculated over the entire set of trajectories: $J_{\text{form}}(\Xi, \xi_a)$.

We then compose the overall cost function as a linear combination between each component, with relative weights λ. The solution Ξ^* is then tracked by each UAV:

$$\begin{aligned} J(\Xi) &= \begin{bmatrix} 1 & \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix} \begin{bmatrix} J_{\text{smooth}}(\Xi) \\ J_{\text{occlusion}}(\Xi) \\ J_{\text{obstacle}}(\Xi) \\ J_{\text{form}}(\Xi) \end{bmatrix} \\ \Xi^* &= \arg\min \quad J(\Xi) \end{aligned} \tag{2}$$

IV. Approach

We now detail the methods we use for camera coordination in the multi-UAV system. As displayed in Equation 1, our overall objective function involves the minimization of 4 sub-objectives, which may often conflict with one another. Our goal is to formulate an algorithm that works in real time in unscripted scenes, and scales to more UAVs without a large computational penalty.

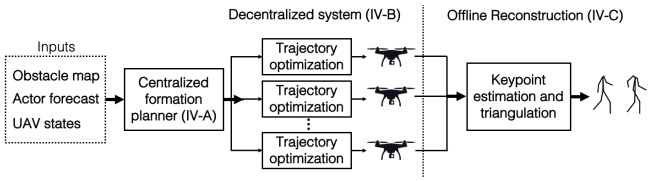


Fig. 2: System overview: a centralized formation planner computes discrete camera positions for optimal target reconstruction. Next, each UAV optimizes smooth trajectories to follow. Finally, the collected multi-view footage and camera poses are used offline to reconstruct a sequence of the target's joints in 3D.

To address the time complexity issue, we break down our method into three main subsystems operate together. First, a centralized motion planner (Sec. IV-A) coordinates desired positions for all cameras simultaneously. Next, on each UAV, a decentralized motion planner (Sec. IV-B) computes the final trajectories for the specific UAV. Finally, an offline skeletal reconstruction module (Sec. IV-C) processes images from all cameras to output the actor's pose vector over time. Fig. 2 depicts the system diagram.

A. Centralized Formation Planning

Our centralized formation planning system parametrizes trajectories as waypoints, i.e. $\xi \in \mathbb{R}^{T\times 3}$, where T is the number of time steps. Actor trajectory ξ_a is forecasted given current actor pose using a Kalman Filter with a constant-velocity model. We assume the UAV heading direction $\psi(t)$ is set to always point the drone from $\xi_{qi}(t)$ towards the actor in $\xi_a(t)$, which can be achieved independently of the aircraft's translation by rotating the UAV's body and camera gimbal. We extend our previous multi-drone cinematography work [24] and plan formation trajectories using a state-space parametrized in spherical coordinates $\{\rho, \theta, \phi\}$ centered on the actor's position (Fig. 3a):

$$\xi_{qi}(t) = \xi_a(t) + \rho \begin{bmatrix} cos(\theta_i)cos(\phi_i) \\ sin(\theta_i)cos(\phi_i) \\ sin(\phi_i) \end{bmatrix} \tag{3}$$

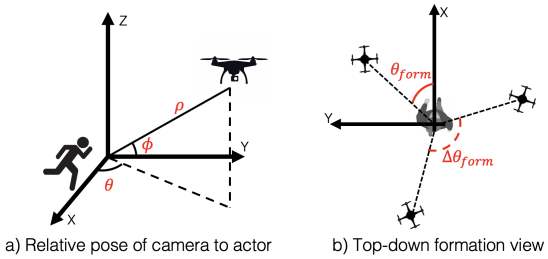


Fig. 3: (a) Spherical camera coordinates relative to actor. (b) Top-down view of formation showing formation yaw θ_{form} and desired yaw angle difference $\Delta\theta_{form}$.

Next, we mathematically formulate the cost functions that the centralized formation planner optimizes for the formation trajectory set Ξ:

i) Formation: To minimize the pose reconstruction error (Eq. 1), the ideal camera formation should maximize the angular distance between the multiple cameras. We define the optimal formation for n UAVs (Fig. 3b) as points with equidistant yaw angles relative to the target, where $\Delta\theta_{\text{form}} = \frac{2\pi}{n}$, and with a special case of $\Delta\theta_{\text{form}} = \frac{\pi}{2}$ for $n = 2$. The desired tilt angle ϕ_{form} and radius ρ_{form} are equal for all UAVs. The cost is calculated as:

$$J_{\text{form}}(\Xi) = \sum_{t=1}^{T}\sum_{i=1}^{n} ||\xi_i(t) - \xi_{i\ \text{form}}(t)|| \tag{4}$$

ii) Safety: To maintain safety, we must reason about the role of obstacles in the environment. First, we transform the environment's occupancy grid into a time-dependent spherical domain centered around the target $\mathcal{G} \rightarrow \mathcal{G}_s^t \in [0, 1]$, as shown in Fig. 4 and Eq. 5.

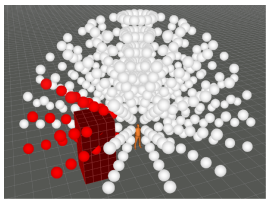


Fig. 4: Visualization of occupancy and occlusion avoidance costs in spherical grid $\mathcal{G}_s^t$, from [24].

$$J_{\text{obstacle}}(\Xi) = \sum_{t=1}^{T}\sum_{i=1}^{n}\int_0^{r_{max}} \mathcal{G}_s^t(\xi_{qi}(t))\ d(\text{volume}) \tag{5}$$

iii) Occlusion avoidance: In order to maintain target visibility at all times, we calculate the occlusion cost as a measure of occupancy along a line $l_i(\tau) = \tau\xi_{qi}(t) + (1-\tau)\xi_a(t)$ between UAV and target:

$$J_{\text{occlusion}}(\Xi) = \sum_{t=1}^{T}\sum_{i=1}^{n}\int_0^{1} \mathcal{G}_s^t(l_i(\tau))\ d\tau \tag{6}$$

Next, we find the optimal sequence of angles for the UAV formation that minimizes the sum of costs. Instead of solving for each UAV path sequentially as in our previous work [24], in this work we assume a fixed drone formation, and only find the optimal yaw angle sequence over T time steps: $\Theta^*_{form} = \{\theta_1, ..., \theta_T\}$.

We define a state space S with all possible formation yaw values, where $|S| = T \times D$, where D is the number of discrete values over the interval $[-\pi, \pi]$. We build a cost map $C : S \rightarrow \mathbb{R}^{|S|}$ that contains the cost of all states, and a cost-to-go map $V : S \rightarrow \mathbb{R}^{|S|}$. In order to make transitions between cells dynamically feasible for the real vehicle, we only allow expansions to neighboring cells in the next ring. Given that we operate in a discrete state-space with a relatively small branching factor and deterministic transitions, a single backwards dynamic programming pass yields the optimal solution in little time. Finally, we build the full formation yaw sequence Θ^*_{form} by selecting neighboring cells with the least cost-to-go at consecutive time steps, starting at the formation's initial yaw θ_0. Algorithm 1 details the process.

B. Decentralized Trajectory Optimization

After calculating the formation angles Θ^*_{form} using the centralized planner, we optimize and smoothen individual

Algorithm 1: Compute formation $\Theta^*_{form} = \{\theta_1, ..., \theta_T\}$

$C \leftarrow J(S)$; ▷ update formation cost map
$V_T \leftarrow C$; ▷ initialize cost-to-go at time T
▷ Begin backwards pass
for $t = T-1, T-2, ..., 1$ **do**
  for $i = 1, ..., D$ **do**
    $V_t^i \leftarrow \arg\min_i V_{t+1}^i + C^i$; ▷ neighbors
  end
end
▷ Begin forward pass
$\Theta^*_{form}(0) = \theta_0$;
for $t = 1, ..., T$ **do**
  $N = \text{neighbors}(\Theta^*_{form}(t-1))$; ▷ connected cells
  $\Theta^*_{form}(t) = \arg\min_N V_t$;
end
return Θ^*_{form}

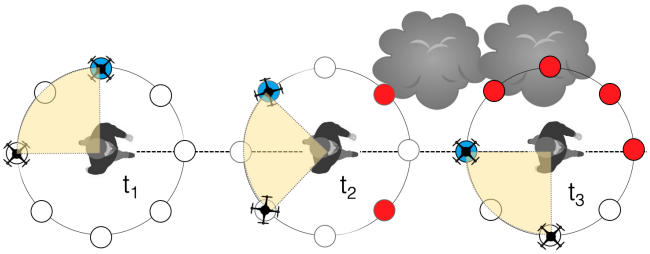


Fig. 5: Centralized formation planning where formation rotates counter-clockwise to avoid trees. Formation cost map is updated for all time steps, with red as high cost. We apply dynamic programming to solve for the full formation yaw sequence Θ^*_{form}, shown in blue.

UAV trajectories at a finer time discretization. This second step is computed on each UAV's local computer using a decentralized planner. While the original waypoints were spaced every 2 seconds over a 10-second horizon, here we achieve finer resolutions with 0.5 s granularity in local planning. We use the local planner described in [12], which uses covariant gradient descent to produce locally optimal trajectories while again considering the costs of smoothness, obstacle and occlusions avoidance, and desired formation position of each UAV individually. In addition, each local planner receives the expected waypoints of all remaining vehicles, and avoids positioning its trajectory within $3m$ of other UAVs. We run the local planner at 5 Hz, and use a PID controller for trajectory tracking at 50 Hz.

C. Offline Skeletal Reconstruction

Once camera images from all UAVs are collected, we post-process the data in an offline phase to generate a sequence of 3D target skeleton poses. We use AlphaPose, a human skeletal keypoint detector [37]–[39], to extract 2D body keypoints from each image. Next, we linearly triangulate each keypoint using each robot's camera pose and image coordinates to obtain for the keypoint's location in world frame.

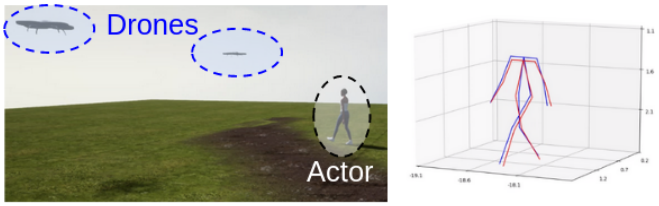


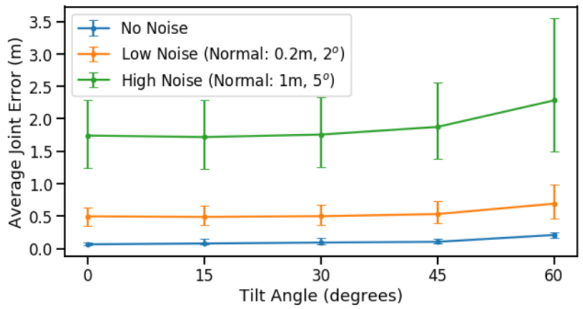
Fig. 6: Experimental Setup: *(Left)* Photo-realistic simulator with an animated character and drone formation. *(Right)* Ground truth skeleton extracted from simulator (blue) and estimated skeleton with our proposed system (red).

V. Experiments

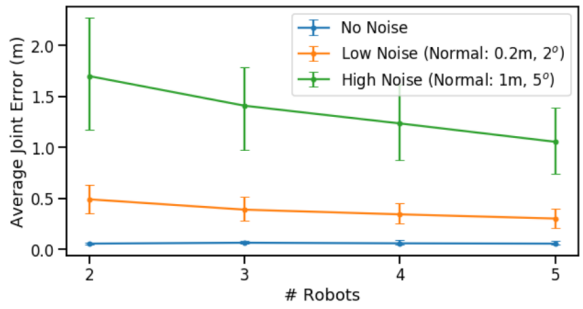
Here we detail the simulated and real-world experiments to validate our multi-UAV 3D motion capture system.

A. Simulation Experiments

Experimental Setup: We quantitatively evaluate our proposed system in a photo-realistic environment, Microsoft AirSim [40] via a custom ROS [41] interface, and directly query ground-truth skeleton points from Unreal Engine. Fig. 6 shows our simulated experimental setup.



(a) Reconstruction Error vs. Formation Tilt Angle



(b) Reconstruction Error vs. # Robots

Fig. 7: *a)* Reconstruction Quality vs. Tilt Angle: We observe better reconstruction and resistance to noise with lower tilt angle, with an angle of 0^o to 30^o giving similar performance. *b)* Reconstruction Quality vs. Number of Robots: With no noise, performance is comparable from $n = 2$ to 5. We observe better resistance to noise as number of robots increases, with marginally decreasing benefit.

Sim E1) Reconstruction quality across tilt angles: Our first experiment's objective is to quantify the benefit of maintaining low formation tilt angle ϕ_{des} for human reconstruction. To do so, we generated 75 seconds of data of an actor walking

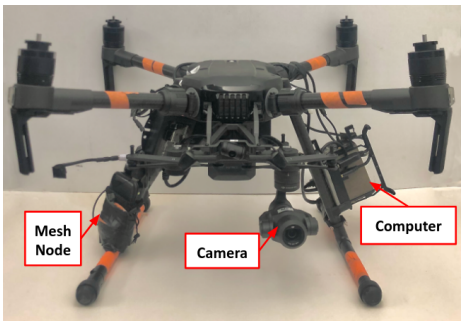


Fig. 8: System hardware: DJI M210 drone, Intel NUC computer, Ubiquiti mesh nodes and Zenmuse X4S camera gimbal.

with two drones at each tilt angle between 0^o to 60^o at increments of 15^o. Fig. 7a shows lower tilt angle yields better reconstruction. This is expected because, firstly at a high tilt angle, most parts of the target's body are occluded and a slight error in image coordinate can result in large 3D reconstruction error. Secondly, it is more likely for the keypoint detector to misidentify limbs at a higher angle, possibly due to lack of training data at such angles. As our camera pose is often noisy for our *in-the-wild* reconstruction application, we also examine how susceptible each tilt angle is to camera pose noise. Within the same noise level, we observe tilt angles of 0^o to 30^o provide comparable reconstruction quality.

Sim E2) Reconstruction quality using more robots: Next, we examine the marginal benefits that more robots bring to reconstruction accuracy. We record 75 seconds of a target walking with with $n = [2, 3, 4, 5]$ drones at a formation tilt angle $\phi_{des} = 15^o$. Figure 7b shows that with no noise, average error is near 0 for all configurations, with a slight decrease at 5 drones. As expected, increasing the number of simultaneous viewpoints helps significantly with noisy camera poses, with decreasing marginal benefits as number of drones increase. At high noise level, the 5-drone configuration reduces error by $\sim 30\%$ from the two-drone setup.

B. Real-World Experiments

Experimental Setup: For real world experiments, we used two DJI M210 drones, one shown in Figure 8. We subsequently refer to these as *drone 1* and *drone 2*. All processing is done onboard an Intel NUC with 8GB of RAM and an Intel Core i7-8550U processor. Drones communicate with each other with a Ubiquiti WiFi mesh access point [42] via the Data Distribution Service networking middleware [43]. The leader drone (*drone 1*) runs the centralized planner and sends estimated actor odometry and formation trajectory to *drone 2* for local decentralized planning. Both drones share current odometry and final optimized trajectory for safety.

An independently controlled DJI Zenmuse X4S gimbal camera records footage. Video frames are processed using an off-the-shelf Intel's OpenVINO MobileNetV2-like pedestrian detector for actor detection.

Our centralized formation planner solves a $10s$ horizon plan

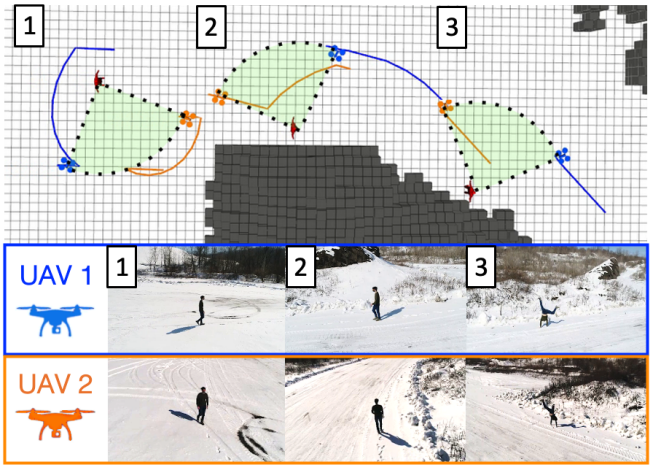


Fig. 9: Real-life flight among obstacle. Our adaptive formation rotates clockwise avoiding the mound to maintain 90^o from each other and a low tilt angle to actor for optimal reconstruction.

with $D = 8$ possible discrete formation yaws. The planner runs at 10Hz with a computation time of $\sim 1.3ms$.

We conduct real-world experiments in a pre-mapped outdoors test site. Range is at $\rho_{des} = 10m$ for safety and formation tilt is set at $\phi_{des} = 15^o$, which from our simulated results renders good reconstruction.

Real E1) Formation obstacle avoidance: We tested the system by recording a moving actor with two drones in a pre-mapped environment. Fig. 9 shows an example trial of our proposed system where the two-drone formation rotated clockwise to avoid colliding with the mound. The drones are therefore able to maintain a low viewing angle while keeping safe. While the central planner runs at 10Hz, we show three representative timesteps of the experiment for clarity. The central planner's output at keyframes for *drones 1* and *2* are colored in blue and orange respectively. Each UAV then optimizes the coarse formation path with its own local planner for a final smooth, obstacle-free trajectory.

Real E2) Adaptive versus fixed formation: For the same initial formation angle, we compare the reconstruction performance with and without our adaptive formation planning. In the *fixed* trial, the two drones go upward and deviate from the desired tilt angle to avoid mound, resulting in a tilt angle of $\sim 60^o$. The high tilt angle results in highly inconsistent reconstruction, due to the increased likelihood of keypoint detection error with examples circled in Fig. 10. Our proposed adaptive formation planning keeps drones at a low tilt angle, significantly improving the reconstruction quality while avoiding obstacles.

Real E3) Reconstructing highly dynamic targets: We evaluate the robustness of our proposed system by reconstructing an actor performing abrupt motion changes and highly dynamic movement: jogging and playing soccer. Figure 11 shows the reconstruction of an actor playing soccer. As seen in the supplementary video, both UAVs were

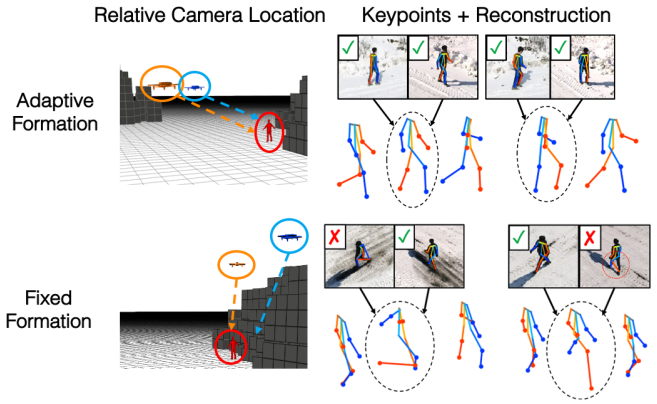


Fig. 10: Reconstruction comparison with and without adaptive formation planning. The formation planning keeps camera at a low tilt angle and significantly improves reconstruction. Without formation planning, the UAVs goes upward to avoid mound. The skeletal keypoint detection fails often for footage collected with fixed formation, resulting in poor reconstruction.

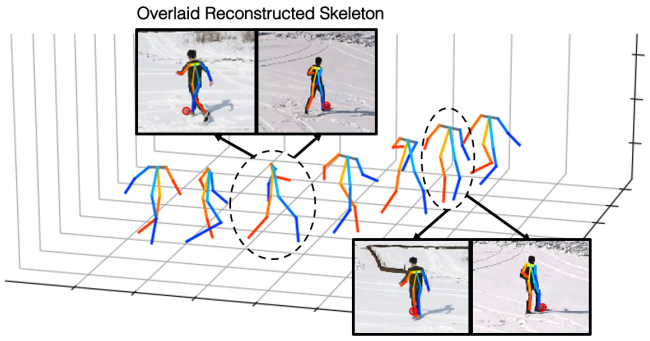


Fig. 11: 3D reconstruction of a highly dynamic real-life actor playing soccer. Our system is able to keep up with abrupt motion changes and provide good reconstruction. Inlet figures show the reprojection of reconstructed 3D skeleton overlaid onto UAV images.

able to maintain view of the target while keeping desired reconstruction formation angle. Inlet images in Figure 11 show the reprojection of the target joints on the UAV's images, with reprojected joints close to actual position.

VI. Conclusion and Discussion

In this paper, we present a collaborative multi-UAV system for 3D human pose reconstruction *in the wild*. First, we develop a multi-camera coordination scheme to maximize 3D reconstruction quality of dynamic targets while avoiding obstacles and occlusions. Our approach consists of two steps: 1) a centralized formation planner to compute best camera formation and 2) a decentralized trajectory optimizer to calculate smooth trajectories. We validate our system in simulated and real-world experiments, and show that it successfully reconstructs targets performing dynamic activities, such as jogging and playing soccer. Additionally, we provide insights into how reconstruction quality changes with our system under different operating conditions, such as number of drones, camera pose uncertainty and tilt angles.

We find multiple directions for future work in our multi-UAV system. We are actively working towards the goal of capturing high-fidelity 4D reconstruction of groups of actors and animals in their natural settings, with research thrusts in onboard multi-actor detection and tracking [44], [45], adaptive multi-agent role reconfiguration to maintain visibility of actors when group splits [46], and human mesh reconstruction [47].

Acknowledgment

The authors thank Arthur Bucker, Andrew Ashley and Sam Triest for their help in field experiments. CH is supported by the Croucher Foundation. This work is supported by the National Science Foundation under grant no. 2024173.

References

[1] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 559–568.

[2] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. Ieee, 2011, pp. 963–968.

[3] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[4] J. C. Chan, H. Leung, J. K. Tang, and T. Komura, "A virtual reality dance training system using motion capture technology," *IEEE Transactions on Learning Technologies*, vol. 4, no. 2, pp. 187–195, 2011.

[5] A. Fern'ndez-Baena, A. Susín, and X. Lligadas, "Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments," in *2012 fourth international conference on intelligent networking and collaborative systems*. IEEE, 2012, pp. 656–661.

[6] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1746–1753.

[7] I. Kitahara, H. Saito, S. Akimichi, T. Ono, Y. Ohta, and T. Kanade, "Large-scale virtualized reality," *Computer Vision and Pattern Recognition, Technical Sketches*, 2001.

[8] R. Collins and T. Kanade, "Multi-camera tracking and visualization for surveillance and sports," *Proceedings of the Fourth International Workshop on Cooperative Distributed Vision*, 2001.

[9] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[10] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.

[11] Y. Zhang and H. S. Park, "Multiview supervision by registration," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 409–417.

[12] R. Bonatti, W. Wang, C. Ho, A. Ahuja, M. Gschwindt, E. Camci, E. Kayacan, S. Choudhury, and S. Scherer, "Autonomous aerial cinematography in unstructured environments with learned artistic decision-making," *Journal of Field Robotics*, 2020.

[13] R. Bonatti, C. Ho, W. Wang, S. Choudhury, and S. Scherer, "Towards a robust aerial cinematography platform: Localizing and tracking moving targets in unstructured environments," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[14] M. Gschwindt, E. Camci, R. Bonatti, W. Wang, and S. Scherer, "Can a robot become a movie director? learning artistic principles for aerial cinematography," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[15] R. Bonatti, Y. Zhang, S. Choudhury, W. Wang, and S. Scherer, "Autonomous drone cinematographer: Using artistic principles to create smooth, safe, occlusion-free trajectories for aerial filming," *International Symposium on Experimental Robotics*, 2018.

[16] C. Gebhardt, B. Hepp, T. Nägeli, S. Stevšić, and O. Hilliges, "Airways: Optimization-based planning of quadrotor trajectories according to high-level user goals," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 2508–2519.

[17] C. Gebhardt, S. Stevsic, and O. Hilliges, "Optimizing for aesthetically pleasing quadrotor camera motion," *ACM Transactions on Graphics*, 2018.

[18] N. Joubert, M. Roberts, A. Truong, F. Berthouzoz, and P. Hanrahan, "An interactive tool for designing quadrotor camera shots," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 238, 2015.

[19] M. Roberts and P. Hanrahan, "Generating dynamically feasible trajectories for quadrotor cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 61, 2016.

[20] T. Nägeli, S. Oberholzer, S. Plüss, J. Alonso-Mora, and O. Hilliges, "Flycon: Real-time environment-independent multi-view human pose estimation with aerial vehicles," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 182:1–182:14, Dec. 2018.

[21] R. Tallamraju, N. Saini, E. Bonetto, M. Pabst, Y. T. Liu, M. Black, and A. Ahmad, "Aircaprl: Autonomous aerial human motion capture using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6678 – 6685, Oct. 2020.

[22] N. Saini, E. Price, R. Tallamraju, R. Enficiaud, R. Ludwig, I. Martinovic, A. Ahmad, and M. J. Black, "Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] R. Tallamraju, E. Price, R. Ludwig, K. Karlapalem, H. H. Bülthoff, M. J. Black, and A. Ahmad, "Active perception based formation control for multiple aerial vehicles," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4491–4498, 2019.

[24] A. Bucker, R. Bonatti, and S. Scherer, "Do you see what i see? coordinating multiple aerial cameras for robot cinematography," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[25] "Skydio," https://www.skydio.com/technology, August 2020.

[26] "Dji mavic," https://www.dji.com/br/mavic, August 2020.

[27] T. Nägeli, L. Meier, A. Domahidi, J. Alonso-Mora, and O. Hilliges, "Real-time planning for automated multi-view drone cinematography," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 132, 2017.

[28] A. Torres-González, A. Alcántara, V. Sampaio, J. Capitán, B. Guerreiro, R. Cunha, and A. Ollero, "Distributed mission execution for aerial cinematography with multiple drones," 2019.

[29] A. Alcántara, J. Capitán, A. Torres-González, R. Cunha, and A. Ollero, "Autonomous execution of cinematographic shots with multiple drones," *IEEE Access*, vol. 8, pp. 201 300–201 316, 2020.

[30] A. Alcántara, J. Capitán, R. Cunha, and A. Ollero, "Optimal trajectory planning for cinematography with multiple unmanned aerial vehicles," *Robotics and Autonomous Systems*, vol. 140, p. 103778, 2021.

[31] L.-E. Caraballo, Ángel Montes-Romero, J.-M. Díaz-Báñez, J. Capitán, A. Torres-González, and A. Ollero, "Autonomous planning for multiple aerial cinematographers," 2020.

[32] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type constraints in uav cinematography for autonomous target tracking," *Inf. Sci.*, vol. 506, pp. 273–294, 2020.

[33] A. Pirinen, E. Gartner, and C. Sminchisescu, "Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[34] X. Zhou, S. Liu, G. Pavlakos, V. Kumar, and K. Daniilidis, "Human motion capture using a drone," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2027–2033.

[35] S. Kiciroglu, H. Rhodin, S. N. Sinha, M. Salzmann, and P. Fua, "Activemocap: Optimized viewpoint selection for active human motion capture," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 100–109.

[36] L. Xu, Y. Liu, W. Cheng, K. Guo, G. Zhou, Q. Dai, and L. Fang, "Flycap: Markerless motion capture using multiple autonomous flying cameras," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 8, pp. 2284–2297, 2018.

[37] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[38] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," *arXiv preprint arXiv:1812.00324*, 2018.

[39] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *BMVC*, 2018.

[40] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*. Springer, 2018, pp. 621–635.

[41] ROS. (2018) Robot operating system (ros). [Online]. Available: http://www.ros.org/

[42] "Unifi mesh access point," February 2020. [Online]. Available: https://store.ui.com/collections/unifi-network-access-points/products/unifi-ac-mesh-ap

[43] G. Pardo-Castellote, "Omg data-distribution service: Architectural overview," in *23rd International Conference on Distributed Computing Systems Workshops, 2003. Proceedings.* IEEE, 2003, pp. 200–206.

[44] Y. Wang, K. Kitani, and X. Weng, "Joint Object Detection and Multi-Object Tracking with Graph Neural Networks," *arXiv:2006.13164*, 2020.

[45] X. Weng, Y. Yuan, and K. Kitani, "End-to-End 3D Multi-Object Tracking and Trajectory Forecasting," *ECCVW*, 2020.

[46] S. Engin and V. Isler, "Active localization of multiple targets using noisy relative measurements," 2020.

[47] Y. Jafarian and H. S. Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *CVPR*, 2021.