AirCap – Aerial Outdoor Motion Capture

Aamir Ahmad[1], Eric Price[1], Rahul Tallamraju[1,2], Nitin Saini[1], Guilherme Lawless[3], Roman Ludwig[1], Igor Martinovic[1], Heinrich H. Bülthoff[4] and Michael J. Black[1]

***Abstract*— This paper presents an overview of the Grassroots project *Aerial Outdoor Motion Capture (AirCap)* running at the Max Planck Institute for Intelligent Systems. AirCap's goal is to achieve markerless, unconstrained, human motion capture (mocap) in unknown and unstructured outdoor environments. To that end, we have developed an autonomous flying motion capture system using a team of aerial vehicles (MAVs) with only on-board, monocular RGB cameras. We have conducted several real robot experiments involving up to 3 aerial vehicles autonomously tracking and following a person in several challenging scenarios using our approach of active cooperative perception developed in AirCap. Using the images captured by these robots during the experiments, we have demonstrated a successful offline body pose and shape estimation with sufficiently high accuracy. Overall, we have demonstrated the first fully autonomous flying motion capture system involving multiple robots for outdoor scenarios.**

Fig. 1: 3D markerless motion capture from fully autonomous micro aerial vehicles (MAVs) with on-board cameras. Multi-exposure image shows the trajectory of the MAVs and the 3D body pose and shape projected onto an image frame from an external camera.

I. Introduction

Human pose and shape estimation, or motion capture (mocap), in outdoor, unstructured environments is a highly relevant and challenging problem. Its wide range of applications include search and rescue [1], coordinating outdoor sports events [2] and facilitating animal conservation efforts in the wild [3]. In indoor settings, similar applications usually make use of body-mounted sensors, artificial markers and static cameras. While such markers might still be usable in outdoor scenarios, dynamic ambient lighting conditions and the impossibility of having environment-fixed cameras make the overall problem difficult. On the other hand, body-mounted sensors are not suitable for some kinds of subjects (e.g., animals in the wild or large crowds of people). Therefore, to address all of these issues, our solution involves a team of micro aerial vehicles (MAVs), tracking subjects by using only on-board monocular cameras and computational units, without any subject-fixed sensor or marker.

Our method consists of a robotic front-end and an optimization-based back-end. The front-end consists of a team of micro aerial vehicles (MAVs), autonomously detecting, tracking and following a person. It is responsible for the online task, which is to continuously estimate the 3D global position of the person and keep him/her centered in the field of view of their on-board camera, while he/she performs activities such as walking, running, jumping, etc. The back-end performs the offline task of human pose and shape estimation using only the acquired RGB images and the MAV's self-localization poses (the camera extrinsics).

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany.
[2] Agents and Applied Robotics Group, IIIT Hyderabad, India.
[3] ISR, Instituto Superior Técnico, Lisbon, Portugal.
[4] Max Planck Institute for Biological Cybernetics, Tübingen, Germany.
Corresponding author email:rahul.tallamraju@tuebingen.mpg.de.

II. AirCap System Overview

Fig. 2 shows the overview of our mocap system. Step 1 in Fig. 2 depicts the robotic font-end which is used to acquire images and save them on-board during a mocap session. Fig. 1 shows one such mocap session using our front-end. Step 2–4 in Fig. 2 form the back-end which is responsible for offline pose and shape estimation using the acquired images.

A. Front-end: Online Data Acquisition

1) System: Fig. 1 shows our MAV-based mocap front-end tracking and following a person. It consists of a team of self-designed 8-rotor MAVs (see in Step 3 in Fig. 2 inset). Each MAV is equipped with a 2MP HD camera, a computer with an Intel i7 processor, an NVIDIA Jetson TX1 embedded GPU and an OpenPilot Revolution flight controller board. We use the flight controller's position and yaw controller as well as its GPS and IMU-based self-pose (position and orientation) estimation functionalities.

2) Detection and Tracking: In order to detect and cooperatively track the person, we use the approach developed in our previous work [4]. Each copter runs a single shot detector (SSD) multibox on the images acquired by its own camera using its on-board GPU to detect the person's outer bounding box on the image frames. A detection rate of $\sim$4 Hz is achieved during the online acquisition. The MAVs then share the person's 2D image bounding box positions and their own 3D self-pose estimates wirelessly between each other. Subsequently, using a cooperative detection and tracking (CDT) filter that runs on-board each MAV's CPU,

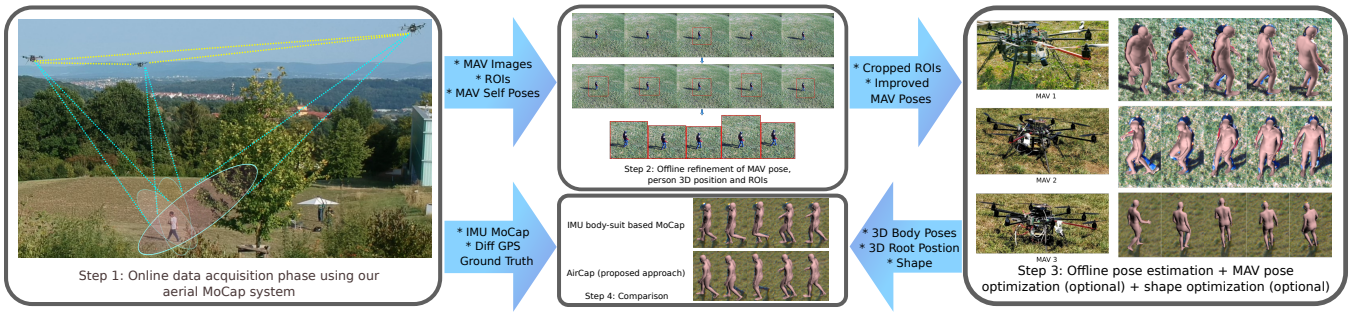

Fig. 2: AirCap System Overview. Step 1 is the online phase while Step 2–4 are parts of the offline phase, as described in the text previously.

they estimate the 3D position of the person's center of mass in a consistent world frame (GPS-frame). Using this method, the MAVs also improve their own 3D self-pose localization. One key feature of the CDT filter is that it allows the detector to focus on the most informative region of image (ROI) on future image frames, thereby making it computationally efficient. Note that even though the detections are obtained at $\sim$4 Hz, the CDT filter runs at $\sim$30 Hz, alternating between the standard prediction and update steps, except that the updates happen at a lower frequency.

3) Navigation and Formation Control: In order to continuously keep the person in the camera FOV of all MAVs, they need to follow him or her. To this end, we developed an active perception-driven MPC-based formation controller [5]. An instance of this controller runs on each MAV to generate waypoints for its own navigation such that it actively minimizes the joint uncertainty in the tracked person's 3D position estimate obtained from CDT (described in the previous sub-section). We formulate it as a locally convex MPC. We do so by decoupling joint uncertainty minimization into i) a convex quadratic objective that maintains a threshold distance to the tracked person, and ii) constraints that enforce angular configurations of the MAVs with respect to (w.r.t.) the person. We derive this decoupling based on Gaussian observation model assumptions used within the CDT algorithm. To guarantee the safety of the motion plans, we incorporate collision avoidance constraints w.r.t. i) other MAVs, ii) the tracked person and iii) static obstacles, only as locally convex constraints. Collision avoidance and angular configuration constraints are inherently non-convex. We preserve convexity in our MPC formulation by converting them to external control input terms embedded inside the MPC dynamics, which are explicitly computed at every iteration of the MPC.

B. Back-end: Offline Pose and Shape Estimation

1) 2D region of interest and MAV offline self pose refinement: In this step, we run the CDT algorithm [4] offline to improve the subject's tracked position estimate and each MAV's self pose estimates. SSD Multibox detector is able to run on every frame in Step 2. The CDT filter leverage these every-frame observations to obtain the ROIs for every image and improve the MAV self pose estimates.

2) Offline Pose and Shape Estimation: Our approach to this step [6] relies on 2D joint detections in each camera. Current methods like AlphaPose [7] and OpenPose [8] are quite accurate even with aerial imagery. To fuse these 2D detections into a 3D pose estimate, we formulate an objective function in which we simultaneously solve for body shape, 3D pose, and 3D camera positions. Pose is represented by relative joint rotations of body parts in a kinematic tree. Specifically, we use the 3D SMPL body model [9] to fuse the noisy estimates. SMPL captures the shape of the human body and this constrains the possible solutions. We project the joints of SMPL onto each of the images (using the estimated camera parameters) and compute the error (robustly) between the predictions and the observed 2D detections. Since the 2D pose detections may be noisy, we regularize the 3D fitting with a learned pose prior called Vposer [10]. Vposer is learned from SMPL fits to hours of motion capture data using a variational auto-encoder (VAE). We solve for camera parameters jointly and constrain them to be similar to those estimated by the MAVs. Further details of our approach are presented in [6].

3) Ground Truth Comparison: We obtain ground truth data to evaluate our reconstructions from two different systems, i) a commercially available IMU MoCap system (Xsens) [11] and ii) a pair of differential GPS modules. IMU system is used to obtain ground truth (GT) data for body pose relative to the root joint. For ground truth SMPL parameters, we use a state of the art IMU MoCap method Sparse Inertial Poser (SIP) [12]. It uses the raw data from Xsens and gives SMPL paramters. However, the global root joint position and orientation from SIP is not reliable for GT comparison. To solve this issue, we use a pair of differential GPS modules, each one attached to a shoulder of the subject to get the position of root joint in global coordinate system. GT of global root orientation still remains unknown as it is not directly measurable with these two GT systems.

III. Results and Demonstration

Figure 4 shows multi-exposure images from 4 mocap experiments. Fig. 4(a) showcases the results based on our previous work [4]. Notice that the MAVs are close to the target person and never uniformly spread around the person's position. Fig. 4(b) shows the results based on our previous work [13]. Notice that the MAVs 2 and 3 are quite close to each other and the resulting formation is non-optimal for uncertainty minimization. In Fig. 4(c), (d) and the image in Fig. 1 the results of our latest approach [5] are shown. Notice that the MAVs are almost uniformly spread around the person's position and maintain an angular configuration with a difference of approx. $\frac{2\pi}{3}$ w.r.t. each other. Moreover,

Fig. 3: Pose and shape estimation results of our approach overlaid on some of the image sequences from one of the MAV's camera. (Left) A walking sequence. (Right) A sequence with arbitrary hand and leg movement.

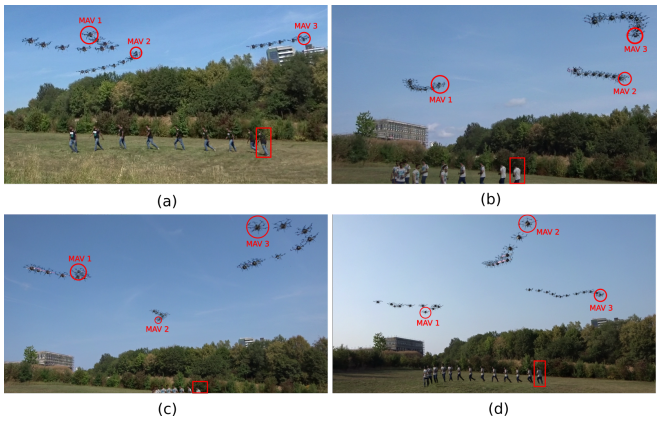

Fig. 4: Multi-exposure images of short sequences from different mocap experiments.

the MAVs successfully maintain the desired safe distance and altitude from the person. We refer the reader to [5] for complete description of experimental setup and results

In Fig. 3 we present qualitative results of our mocap back-end, the offline pose and shape estimation method [6]. The estimated SMPL body pose and shapes are overlaid on images obtained from the online data collection step. It showcases i) a walking sequence where the human subject is randomly walking in different directions and ii) a sequence where the human subject is performing arbitrary hand and leg movements while standing or jumping in the same location. In [6], a quantitative comparison with ground truth (GT) motion capture of the person is also presented. For GT, an IMU and differential GPS-based body suit was used.

In [5] and [6] further experimental results are presented. They include various ablation studies, comparisons on how multiple robots improve pose and shape estimates, etc.

IV. Conclusion and Future Work

We presented the first successful demonstration of full-body markerless motion capture from autonomous flying vehicles. AirCap addresses the challenges of i) online image data acquisition of a tracked human subject by multiple fully autonomous MAVs, and ii) human body pose and shape estimation using the acquired image dataset. Our first contribution consisted of a cooperative detection and tracking algorithm that actively selects image regions of interests such that CNN-based detectors can run on-board and in realtime on our flying robots. We then introduced a decentralized convex MPC-based algorithm for the MAVs to actively track and follow a moving person in outdoor environments and in the presence of static and dynamic obstacles. Finally, we show how we leverage state-of-the-art 2D human joint detection methods as noisy sensors and fuse them to obtain consistent 3D estimates of human pose and shape. One of the most important advantages of our method is that it completely removes the need for a subject preparation step, therefore, allowing in-the-wild motion capture of any type of subject. Extending our method to larger and complex outdoor scenarios as well as to multiple human and animal subjects are our next steps.

V. Acknowledgments

We are grateful to Mason Landry and Marcin Odelga for experiment assistance and to all the reviewers and editors of our works [4], [5], [13] and [6].

References

[1] M. Andriluka, P. Schnitzspan, J. Meyer, S. Kohlbrecher, K. Petersen, O. von Stryk, S. Roth, and B. Schiele, "Vision based victim detection from unmanned aerial vehicles," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2010, pp. 1740–1747.

[2] "MULTIDRONE: H2020-ICT-2016-2017 H2020-ICT-2016-1 - Robotics and Artificial Intelligence Project," 2008. [Online]. Available: https://multidrone.eu/multidrone-in-short/

[3] G. A. M. d. Santos, Z. Barnes, E. Lo, B. Ritoper, L. Nishizaki, X. Tejeda, A. Ke, H. Lin, C. Schurgers, A. Lin, and R. Kastner, "Small unmanned aerial vehicle system for wildlife radio collar tracking," in *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*, Oct 2014, pp. 761–766.

[4] E. Price, G. Lawless, R. Ludwig, I. Martinovic, H. H. Bülthoff, M. J. Black, and A. Ahmad, "Deep neural network-based cooperative visual tracking through multiple micro aerial vehicles," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3193–3200, Oct 2018.

[5] 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*, 2019.

[6] N. Saini, E. Price, R. Tallamraju, R. Enficiaud, R. Ludwig, I. Martinovi, A. Ahmad, and M. Black, "Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles," in *International Conference on Computer Vision*, Oct. 2019.

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

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[10] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *CVPR*, 2019.

[11] "Xsens motion capture system," https://www.xsens.com.

[12] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 349–360, May 2017. [Online]. Available: https://doi.org/10.1111/cgf.13131

[13] R. Tallamraju, S. Rajappa, M. J. Black, K. Karlapalem, and A. Ahmad, "Decentralized mpc based obstacle avoidance for multi-robot target tracking scenarios," in *Proceedings of IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, 2018.