ActiveMoCap: Optimized Drone Flight for Active Human Motion Capture

Sena Kiciroglu¹, Helge Rhodin¹,², Sudipta Sinha³, Mathieu Salzmann¹, and Pascal Fua¹

¹CVLAB, EPFL, Lausanne
²Imager Lab, UBC, Vancouver
³Microsoft Research, Redmond

Abstract

The accuracy of monocular 3D human pose estimation depends on the viewpoint from which the image is captured. While camera-equipped drones provide control over this viewpoint, automatically positioning them at the location which will yield the highest accuracy remains an open problem. This is the problem that we address in this paper. Specifically, given a short video sequence, we introduce an algorithm that predicts the where a drone should go in the future frame so as to maximize 3D human pose estimation accuracy. A key idea underlying our approach is a method to estimate the uncertainty of the 3D body pose estimates. We integrate several sources of uncertainty, originating from a deep learning based regressors and temporal smoothness. The resulting motion planner leads to improved 3D body pose estimates and outperforms or matches existing planners that are based on person following and orbiting.

1. Introduction

Monocular solutions to 3D human pose estimation have become increasingly competent in recent years, but their accuracy remains relatively low. In this paper, we explore the use of a moving camera whose motion we can control to resolve ambiguities inherent to monocular 3D reconstruction and to increase accuracy. This is known as active vision and has received surprisingly little attention in the context of using modern approaches to body pose estimation. Fig. 1 depicts our approach for a camera carried by a drone, a key use-case for it.

In this paper, we introduce an algorithm designed to continuously position a moving camera, possibly mounted on a drone, at optimal locations to maximize the 3D pose estimation accuracy for a freely moving subject. We achieve this by moving the camera in 6D pose space to locations that maximize a utility function designed to predict reconstruction accuracy. This is non-trivial because reconstruction accuracy cannot be directly used in our utility function. Estimating it while planning the motion would require knowing the true person and drone position, a chicken and egg problem. Instead we use prediction uncertainty as a surrogate for accuracy. This is a common strategy to handle rigid scenes in robotics, for example to send the robot to locations where its internal map is most incomplete [18]. However, in our situation, estimating uncertainty is much more difficult since multiple sources of uncertainty need to be considered. They include uncertainties about what the subject will do next, the reliability of the pose estimation algorithm, and the accuracy of distance estimation along the camera’s line of sight.
Our key contribution is therefore a formal model that provides an estimate of the posterior variance and probabilistically fuses these sources of uncertainty with appropriate prior distributions. This has enabled us to develop an active motion capture technique that takes raw video footage as input from a moving aerial camera and continuously computes future target locations for positioning the camera, in a way that is optimized for human motion capture. We demonstrate our algorithm in two different scenarios and compare it against standard heuristics, such as constantly rotating around the subject and remaining at a constant angle with respect to the subject. We find that when allowed to choose the next location without constraints, our algorithm outperforms the baselines consistently. For simulated drone flight, our results are on par with constant rotation, which we conclude is the best trajectory to choose in the case of no obstacles blocking the circular flight path.

2. Related work

Most recent approaches to markerless motion capture rely on deep networks that regress 3D pose from single images [14, 15, 19, 34, 23, 29, 20, 37, 32, 31, 36]. While a few try to increase robustness by enforcing temporal consistency [21], none considers the effect that actively controlling the camera may have on accuracy. The methods most closely related to what we do are therefore those that optimize camera placement in multi-camera setups and those that are used to guide robots in a previously-unknown environment.

Optimal Camera Placement for Motion Capture. Optimal camera placement is a well-studied problem in the context of static multi-view setups. Existing solutions rely on maximizing image resolution while minimizing self-occlusion of body parts [3, 11] or target point occlusion and triangulation errors [25]. However, these methods operate offline and on pre-recorded exemplar motions. This makes them unsuitable for motion capture using a single moving camera that films a priori unknown motions in a much larger scene where estimation noise can be high.

Concurrent work [22] optimizes multiple cameras poses for triangulation of joints in a dome environment using a self-supervised reinforcement learning approach. In our case, we consider the monocular problem. Our method is not learning based, we try to obtain the next best view from the loss function itself.

View Planning for Static and People Reconstruction. There has been much robotics work on active reconstruction and view planning. This usually involves moving so as to maximize information gain while minimizing motion cost, for example by a discretizing space into a volumetric grid and counting previously unseen voxels [13, 17] or by accumulating estimation uncertainty [18]. When a coarse scene model is available, an optimal trajectory can be found using offline optimization [28, 12]. This has also been done to achieve desired aesthetic properties in cinematography [10].

Another approach is to use reinforcement learning to define policies [6] or to learn a metric [11] for later online path planning. These methods deal with rigid unchanging scenes, except the one in [5] that performs volumetric scanning of people during information gain maximization. However, this approach can only deal with very slowly moving people who stay where they are.

Human Motion Capture on Drones. Drones can be viewed as flying cameras and are therefore natural targets for our approach. One problem, however, is that the drone must keep the person in its field of view. To achieve this, the algorithm of [38] uses 2D human pose estimation in a monocular video and non-rigid structure from motion to reconstruc the articulated 3D pose of a subject, while that of [16] reacts online to the subject’s motion to keep them in view and to optimize for screen-space framing objectives. In [17], this was integrated into an autonomous system that actively directs a swarm of drones and simultaneously reconstructs 3D human and drone poses from onboard cameras. This strategy implements a pre-defined policy to stay at constant distance to the subject and uses pre-defined view angles (90° between two drones) to maximize triangulation accuracy. This enables mobile large-scale motion capture, but relies on markers for accurate 2D pose estimation. In [45], three drones are used for markerless motion capture, using an RGBD video input for tracking the subject. In short, existing methods either optimize for drone placement but for mostly rigid scenes, or estimate 3D human pose but without optimizing the camera placement. [22] performs optimal camera placement for multiple cameras. Here, we propose an approach that aims to find the best next drone location for monocular view so as to maximize 3D human pose estimation accuracy.

3. Active Human Motion Capture

Our goal is to continually pre-position the camera in 6D pose space so that the images it acquires can be used to achieve the best overall human pose estimation accuracy. What makes this problem challenging is that, when we decide where to send the camera, we do not yet know where the subject will be and in what position exactly. We therefore have to guess. To this end, we propose the following three-step approach depicted by Fig. 1:

1. Estimate the 3D human pose up to the current time instant.
2. Predict the person’s location and pose by the time the
camera acquires the next image, including an uncertainty estimate.

3. Cause the camera to assume the optimal 6D pose given that estimate. \( \mathbf{v} \)

We will consider two ways the camera can move. In the first, the camera can teleport from one location to the next without restriction. This can be simulated using a multi-camera setup and allows us to explore the theoretical limits of this approach. In the second, more realistic scenario, the camera is carried by a drone, and we must take into account physical limits about the motion it can undertake.

### 3.1. 3D Pose Estimation

We will consider two ways the camera can move. In the first, the camera setup and allows us to explore the theoretical limits of this approach. In the second, more realistic scenario, the camera is carried by a drone, and we must take into account physical limits about the motion it can undertake.

#### 3.1.1 Formulation

Our primary goal is to improve the global 3D human pose estimation of a subject changing position and pose. We optimize the time-varying pose trajectories across the last \( k \) frames. Let \( t \) be the last observed frame. We capture the trajectory of poses \( \Theta^t \) to \( \Theta^t \) in the pose matrix \( \Theta \). We then write an energy function

\[
E_{\text{pose}} = E_{\text{proj}}(\Theta, M, D) + E_{\text{lift}}(\Theta, L) + E_{\text{smooth}}(\Theta) + E_{\text{bone}}(\Theta, b) .
\]

The individual terms are defined as follows. The lift term, \( E_{\text{lift}} \), leverages the 3D pose estimates, \( L \), from LiftNet \cite{32}. Because these are relative to the hip and without absolute scale, we subtract the hip position from our absolute 3D pose, \( \Theta^t \), and apply a scale factor \( \rho \) to \( L \) to match the bone lengths \( b_{\text{calib}} \) in the least-square sense. We write

\[
E_{\text{lift}}(\Theta, L) = \omega_l \sum_{i=t-k}^{t} \| m \cdot L^i - (\Theta^i - \Theta^i_{\text{hip joint}}) \|_2^2 ,
\]

with \( \omega_l \) its relative weight.

The projection term measures the difference between the detected 2D joint locations and the projection of the estimated 3D pose in the least-square sense. We write it as

\[
E_{\text{proj}}(\Theta, M, D) = \omega_p \sum_{i=t-k}^{t} \| M^i - \Pi(\Theta^i, D^i, K) \|_2^2 ,
\]

where \( \Pi \) is the perspective projection function, \( K \) is the matrix of camera intrinsic parameters, and \( \omega_p \) is a weight that controls the influence of this term.

The smoothness term exploits that we are using a continuous video feed and that the motion is smooth by penalizing velocity computed by finite differences as

\[
E_{\text{smooth}}(\Theta) = \omega_s \sum_{i=t-k+1}^{t} \| (\Theta^{i+1} - \Theta^i) \|_2^2 .
\]

with \( \omega_s \) as its weight.

To further constrain the solution space, we use our knowledge of the bone lengths \( b_{\text{calib}} \) found during calibration and penalize deviations in length. The length of each bone \( b \) in the set of all bones \( b_{\text{all}} \) is found as \( b^t_b = \| (\Theta_{b_1} - \Theta_{b_2}) \|_2 \) for frame \( t \). The bone length term is then defined as

\[
E_{\text{bone}}(\Theta) = \omega_b \sum_{i=t-k}^{t} \sum_{b \in b_{\text{all}}} d(b^t_b, b_{\text{calib},b}) ,
\]

with \( \omega_b \) as its weight.

The complete energy \( E_{\text{pose}} \) is minimized by gradient descent at the beginning of each control cycle, to get a pose estimate for control. The resulting pose estimate \( \Theta \) is the maximum a posteriori estimate in a probabilistic view.

#### 3.1.2 Calibration Mode

Calibration mode only has to be run once for each subject to find the bone lengths, \( b_{\text{calib}} \). In this mode, the scene is assumed to be stationary. The situation is equivalent to having the scene observed from multiple stationary cameras, such as in \cite{27}. We find the single static pose \( \Theta^c \) that minimizes

\[
E_{\text{calib}} = E_{\text{proj}}(\Theta^c, M, D) + E_{\text{symmetry}}(\Theta^c) .
\]
In this objective, the projection term, \( E_{\text{proj}} \), is akin to the one in our main formulation but acts on all calibration frames. It can be written as

\[
E_{\text{proj}}(\Theta^c, M, D) = \omega_p \sum_{i=0}^{t} \| M^i - \Pi(\Theta^c, D^i, K) \|_2^2,
\]

with \( \omega_p \) controlling its influence. The symmetry term, \( E_{\text{symmetry}} \), ensures that the left and right limbs of the estimated skeleton have the same lengths by penalizing the squared difference between length of the left and right bones.

3.2. Best Next View Selection

Our goal is to find the best next view for the drone at the future time step \( t + 1, D^{t+1} \). We will model the uncertainty of the pose estimate in a probabilistic setting. Let \( p(\Theta|M, D, L, b) \) be the posterior distribution of poses. Then, \( E_{\text{pose}} \) is its negative logarithm and its minimization corresponds to maximum a posteriori (MAP) estimation. In this formalism, the sum of the individual terms in \( E_{\text{pose}} \) models that our posterior distribution is composed of independent likelihood and prior distributions. For a purely quadratic term, \( E(x) = \omega (x - \mu)^2 \), the corresponding distribution \( p_E = \exp(-E) \) is a Gaussian with mean \( \mu \) and standard deviation \( \sigma = \frac{1}{\sqrt{2\omega}} \). Notably, \( \sigma \) is directly linked to the weight \( \omega \) of the energy. Most of our energy terms involve non-linear operations, such as perspective projection in \( E_{\text{proj}} \), and therefore induce non-Gaussian distributions, as visualized in Fig. 2. Nevertheless, as for the simple quadratic case, the weights \( \omega_p \) and \( \omega_l \) of \( E_{\text{proj}} \) and \( E_{\text{lift}} \) can be interpreted as surrogates for the amount of measurement noise in the 2D and 3D pose estimates.

A good measure of uncertainty is the sum of the eigenvalues of the covariance \( \Sigma_p \) of the underlying distribution \( p \). The sum of the eigenvalues captures the spread of a multivariate distribution with a single variable, similarly to the variance in the univariate case. To exploit this uncertainty estimation for our problem, we now extend \( E_{\text{pose}} \) to model not only the current and past poses but also the future ones and condition it on the choice of the future drone position. To determine the best next drone pose, we sample candidate positions and chose the one with the lowest uncertainty. This process is illustrated in Figure 3.

**Future pose forecasting.** In our setting, accounting for the dynamic motion of the person is key to successfully positioning the camera. We model the motion of the person from the current frame \( t \) to the next \( M \) future frames \( t + i, i \in \{1, \ldots, M\} \) linearly, i.e. we aim to keep the velocity of the joints constant across our window of frames. We also constrain the future poses by the bone length term. The future pose vectors \( \Theta^{t+i} \) are constrained by the smoothness and bone length terms, but for now not by any image-based term since the future images are not yet available at time \( t \). Minimizing this extended \( E_{\text{pose}} \) for future poses gives the MAP poses \( \Theta^{t+i} \). It continues the motion \( \Theta^{t-K} \) smoothly while maintaining the bone lengths. As we predict only the near future, we have found this simple extrapolation to be sufficient. While more advanced methods [9] could be applied to forecast further, we leave this as future work.

**Future measurement forecasting.** We aim to find the future drone position, \( D^{t+1} \), that reduces the posterior uncertainty, but we do not have footage from future viewpoints to condition the posterior on. Instead, we use the predicted...
future human pose $\hat{\Theta}^{t+i}, i \in \{1, \ldots, M\}$, as a proxy for $L^{t+i}$ and approximate $M^{t+i}$ with the projection

$$M^{t+1} = \Pi(\hat{\Theta}^{t+1}, D^{t+1}, K).$$

At first glance, constraining the future pose on these virtual estimates in $E_{\text{pose}}$ does not add anything since the terms $E_{\text{proj}}$ and $E_{\text{init}}$ are zero at $\hat{\Theta}^{t+1}$ by this construction. However, it changes the energy landscape and models how strong a future observation would constrain the pose posterior. In particular, the projection term, $E_{\text{proj}}$, narrows down the solution space in the direction of the image plane but cannot constrain it in the depth direction, creating an elliptical uncertainty as visualized in Fig 3. The combined influence of all terms is conveniently modeled as the energy landscape of $E_{\text{pose}}$ and its corresponding posterior.

In our current implementation we assume that the 2D and 3D detections are affected by pose-independent noise, and their variance is captured by $\omega_q$ and $\omega_p$, respectively. These factors could, in principle, be view dependent and in relation to the person’s pose. For instance, [3] may be more accurate at reconstructing a front view than a side view. However, while estimating the uncertainty in deep networks is an active research field [24], predicting the expected uncertainty for an unobserved view has not yet been attempted in the pose estimation literature. It is an interesting avenue for future work.

**Variance estimator.** $E_{\text{pose}}$ and its corresponding posterior has a complex form due to the projection and prior terms. Hence, the sought-after covariance $\Sigma_p$ cannot be expressed in closed form and approximating it by sampling the space of all possible poses would be expensive. Instead, for the sake of uncertainty estimation, we approximate $p(\Theta|D,M,L,b)$ locally with a Gaussian distribution $q$, such that

$$\Sigma_p(\Theta|D,M,L) \approx \Sigma_q where q = N(\Theta|\hat{\Theta}, \Sigma_q),$$

with $\hat{\Theta}$ and $\Sigma_q$ the Gaussians mean and covariance matrix, respectively. Such an approximation is exemplified in Figure 2. For a Gaussian, the covariance of $q$ can be computed in closed form as the inverse of the Hessian of the negative log likelihood, $\Sigma_q = H^{-1}_{-\log q}$, where $H_{-\log q} = \frac{\partial^2 \log q(\Theta)}{\partial \Theta^2} \bigg|_{\Theta = \hat{\Theta}}$. Under the Gaussian assumption, $\Sigma_p$ is thereby well approximated by the second order gradients, $H_{\hat{\Theta}^{t+1}}$, of $E_{\text{pose}}$. Our experiments show that this simplification holds well for all of the introduced error terms, except for the bone length one, which we therefore exclude from uncertainty estimation.

To select the view with minimum uncertainty among a set of $K$ candidate drone positions, we therefore

1. optimize $E_{\text{pose}}$ once to forecast $M$ human poses $\Theta^{t+i}$, for each $1 \leq i \leq M$
2. use these forecasted poses to set $L^{t+i}$ and $M^{t+i}$ for each $1 \leq i \leq M$, for each candidate position $c$,
3. compute the second order derivatives of $E_{\text{pose}}$ for each $c$, which form $H_c$, and
4. compute and sum up the respective eigenvalues to select the candidate with the least uncertainty.

**Discussion.** In principle, $p(\Theta|M,D,L,b)$, i.e. the probability of the most likely pose, could also act as a measure of certainty, as implicitly used in [25] on a known motion trajectory to minimize triangulation error. However, the term $E_{\text{proj}}(\Theta, M)$ of $E_{\text{pose}}$ is zero for the future time step $t + i$, because the projection of $\Theta^{t+i}$ is by construction equal to $M^{t+i}$ and therefore uninformative. Another alternative that has been proposed in the literature is to approximate the covariance through first order estimates [33], as a function of the Jakobi matrix. However, as also the first order gradients of $E_{\text{proj}}$ vanish at the MAP estimate, this approximation is not possible in our case.

### 3.3. Drone Control Policies and Flight Model

We control the flight of our drone by passing it the desired velocity vector and the desired yaw rotation amount with the maximum speed kept constant at 5 m/s. The drone is sent new commands once every $\Delta t = 0.2$ seconds. Modeling the flight of the drone allows us to foresee the possible positions the drone will be able to reach when we give it various commands. Since our future $M$ locations By forecasting the future $M$ locations of the drone, we can predict the 2D pose estimations $\hat{M}^{t+i}$ for each $\{i\}_{i=1}^M$ more accurately.

We model the drone flight in the following manner. We assume that the drone moves with constant acceleration during a time step $\Delta t$. If the drone has current position $x_{\text{current}}$ and velocity $V_{\text{current}}$, then with acceleration $a_{\text{current}}$, its next position $x_{\text{goal}}$ in $\Delta t$ time will be

$$x_{\text{goal}} = x_{\text{current}} + V_{\text{current}} \Delta t + 0.5 a_{\text{current}} \Delta t^2.$$  

We model our input to the system as the acceleration, $a_{\text{input}}$. The direction of the acceleration is assumed to be the direction of the velocity vector we pass as the movement command to the simulator, and the magnitude is a value determined through least-square minimization. The current acceleration $a_{\text{input}}$ at time $t$ is found as a weighted average of the input acceleration and the acceleration of the previous step $a_{\text{previous}}$. This can be written as

$$a_{\text{input}} = \alpha a_{\text{input}} + (1-\alpha) a_{\text{previous}}.$$  

By estimating the future positions of the drone, we are able to forecast more accurate future 2D pose estimations, leading to more accurate decision making. Examples of predicted trajectories are shown in Figure 4. Further details are provided in the supplementary material.
4. Evaluation

In this section we evaluate the improvement on 3D human pose estimation that is achieved through optimization of the drone flight.

Simulation environment. Although [26, 2, 32] run in real time, and online SLAM from a monocular camera [8] is possible, we use a drone simulator since the integration of all components onto constrained drone hardware is difficult and beyond our expertise. We make simulation realistic by driving our characters with real motion capture data from the CMU Graphics Lab Motion Capture Database [39] and using the AirSim [30] drone simulator that builds upon the Unreal game engine and therefore produces realistic images of natural environments. An image of AirSim is shown in Figure 5. Simulation also has the advantage that the same experiment can be repeated with different parameters and be directly compared to baseline methods and ground-truth motion.

Simulated test set. We test our approach on three motions from the CMU database that increase in difficulty: Walking straight (subject 2, trial 1), Dance with twirling (subject 5, trial 8), and Running in a circle (subject 38, trial 3). Additionally, we use a validation set consisting of Basketball dribble (subject 6, trial 13), and Sitting on a stool (subject 13, trial 6), to conduct a grid search for hyperparameters.

Real test set. To show that our planner also works outside the simulator, we evaluate our approach on a section of the MPI-INF-3DHP dataset, which includes motions such as running around in a circle and waving arms in the air. The dataset provides 14 fixed viewpoints that are at varying distances from one another and from the subject, as depicted in Figure 7. In this case, the best next view is restricted to one of the 14 fixed viewpoints. This dataset lets us evaluate whether the object detector of [26], the 2D pose estimation method of [3], and the 3D pose regression technique of [32] are reliable enough in real environments. Since we cannot control the camera in this setting, we remove those cameras from the candidate locations where we predict that the subject will be out of the viewpoint.

Baselines. Existing drone-based pose estimation methods use predefined policies to control the drone position relative to the human. Either the human is followed from a constant angle and the angle is set externally by the user [17] or the drone undergoes a constant rotation around the human [38]. As another baseline, we use a random decision policy, where the drone picks uniformly randomly among the proposed viewpoints. Finally, the oracle is obtained by moving the drone to the viewpoint where the reconstruction in the next time step will have the lowest average error, which is achieved by exhaustively trying all proposed viewpoints with the corresponding image in the next
Hyper parameters. We set the weights of the loss term for the reconstruction as follows: $\omega_p = 0.0001$ (projection), $\omega_s = 1$ (smoothness), $\omega_t = 0.1$ (lift term), $\omega_b = 1$ (bone length), which were found by grid search. We set the weights for the decision making as $\omega_p = 0.001$, $\omega_s = 1$, $\omega_t = 0.1$, $\omega_b = 1$. Our reasoning is, we need to set the weights of the projection and lift terms slightly lower because they are estimated with large noise, which is introduced by the neural networks or as additive noise. However, they do not need to be as low for the uncertainty estimation.

### 4.1. Analyzing Reconstruction Accuracy

We report the mean Euclidean distance per joint in meters in the middle frame of the temporal window we optimize over. In the toy example, the size of the temporal window is set to $k = 2$ past frames and 1 future frame, and for the drone flight simulations, to $k = 6$ for past frames and 3 future frames.

**Simulation Initialization.** The frames are initialized by back-projecting the 2D joint locations estimated in the first frame, $M^t = 0$, to a distance $d$ from the camera that is chosen such that the back-projected bone lengths match with the average human height. We then refine this initialization by running the optimization without the smoothness term, as there is only one frame. All the sequences are evaluated for 120 frames, with the animation sequences played at 5 Hz.

**Toy Example: Simulating Teleportation.** To understand whether our uncertainty predictions for potential viewpoints coincide with the actual 3D pose errors we will have at these locations, we run the following simulation: We sample a total of 18 points on a ring around the person, as shown in Fig. 6 and allow the drone to teleport to any of these points. We optimize over a total of $k = 2$ past frames and forecast 1 frame into the future. The reasoning behind our choice of window size is we wanted to emphasize the importance of the next choice of frame.

We perform two variants of this experiment. In the first one, we simulate the 2D and 3D pose estimates, $M, L$, by adding Gaussian noise to the ground-truth data. The mean and standard deviation of this noise is set as the error of [2] and [32], run on the validation set of animations. Figure 6 shows a comparison between the ground truth values, noisy ground truth values and the network results. The results of this experiment are reported in Table 1 where we also provide the standard deviations across 5 trials with varying noise and starting from different viewpoints. As a second variant, we use [2] and [32] on the simulator images to obtain the 2D and 3D pose estimates. The results are in the supplementary material.

Altogether, the results show that our active motion planner achieves consistently lower error values than the baselines and we come the closest to achieving the best possible error for these sequences and viewpoints, despite having no access to the true error. The random baseline also performs quite well in these experiments, as it takes advantage of the drone teleporting to a varied set of viewpoints. The trajectories generated by our active planner and the baselines is depicted in Figure 9. Importantly, Figure 6 evidences that our predicted uncertainties accurately reflect the true pose errors, thus making them well suited to our goal.

**Simulating Drone Flight.** To evaluate more realistic cases where the drone is actively controlled and constrained to only move to nearby locations, we simulate the drone flight using the AirSim environment. While simulating drone flight, we target a fixed radius of 7m from the subject and therefore provide direction candidates that lead to preserving this distance. We do not provide samples at different distances, as moving closer is unsafe and moving far-
In this set of experiments, we fly the drone using the simulator's realistic physics engine. To this end, we sample 9 candidate directions in the directions up, down, left, right, up-right, up-left, down-right, down-left and center. We then predict the 3 consecutive future locations using our simplified (closed form) physics model, to get and estimate where the drone will be at when continuing in each of the 9 directions. We then estimate the uncertainty at these sampled viewpoints and choose the minimum.

We achieve comparable results to constant rotation on simulated drone flight. In fact, except for the first few frames where the drone starts flying, we observe the same trajectory as constant rotation, only the rotation direction varies. Constant rotation being optimal in this setting is not counter-intuitive, as constant rotation is very useful for preserving momentum. This allows the drone to sample viewpoints as far apart from one another as possible, while keeping the subject in view. Figure 10 depicts the different baseline trajectories and the active trajectory.

5. Limitations and Future Work

Our primary goal was to make uncertainty estimation tractable, further improvements are needed to run it on an embedded drone system. The current implementation runs at 0.1Hz, however, the optimization is implemented in Python using the convenient but slow automatic differentiation of Pytorch to obtain second derivatives.

Currently, we assume a constant error for the 2D and 3D pose estimates. We will investigate in future work as how to derive situation dependent noise models of deep neural networks.

We have considered a physically plausible drone model but neglected physical obstacles and virtual no-go areas that would restrict the possible flight trajectories. In the case of complex scenes with dynamic obstacles, we expect our algorithm to outperform any simple, predefined policy.

6. Conclusion

We have proposed a theoretical framework for estimating the uncertainty of future measurements from a drone. This permits us to improve 3D human pose estimation by optimizing the drone flight to visit those locations with the lowest expected uncertainty. We have demonstrated with increasingly complex examples, in simulation with synthetic and real footage, that this theory translates to closed-loop drone control and improves pose estimation accuracy. Key to the success of our approach is the integration of several sources of uncertainty. In future work, we would like to find new ways of estimating the uncertainty of the deployed deep learning methods and extend our work to optimize drone trajectories for different computer vision tasks.
References

[1] A. Aissaoui, A. Ouafi, P. Pudlo, C. Gille, Z.-E. Baarir, and A. Taleb-Ahmed. Designing a Camera Placement Assistance System for Human Motion Capture Based on a Guided Genetic Algorithm. Virtual reality, 22(1):13–23, 2018.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Conference on Computer Vision and Pattern Recognition, pages 1302–1310, 2017.

[3] Y. Chao, J. Yang, B. Price, S. Cohen, and J. Deng. Forecasting Human Dynamics from Static Images. In Conference on Computer Vision and Pattern Recognition, 2017.

[4] X. Chen and J. Davis. Camera Placement Considering Occlusion for Robust Motion Capture. Computer Graphics Laboratory, Stanford University, Tech. Rep., 2(2.2):2, 2000.

[5] W. Cheng, L. Xu, L. Han, Y. Guo, and L. Fang. ihuman3d: Intelligent human body 3d reconstruction using a single flying camera. In 2018 ACM Multimedia Conference on Multimedia Conference, pages 1733–1741. ACM, 2018.

[6] S. Choudhury, A. K. G., Ranade, and D. Dey. Learning to gather information via imitation. In ICRA, 2017.

[7] J. Daudelin and M. Campbell. An adaptable, probabilistic, next-best view algorithm for reconstruction of unknown 3-d objects. IEEE Robotics and Automation Letters, 2(3):1540–1547, 2017.

[8] A. J. Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-Time Single Camera Slam. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1052–1067, June 2007.

[9] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent Network Models for Human Dynamics. In International Conference on Computer Vision, 2015.

[10] C. Gebhardt, S. Stevsic, and O. Hilliges. Optimizing for Aesthetically Pleasing Quadrotor Camera Motion. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH), 37(4):90:1–90:11, 2018.

[11] B. Hepp, D. Dey, S. Sinha, A. Kapoor, N. Joshi, and O. Hilliges. Learn-To-Score: Efficient 3D Scene Exploration by Predicting View Utility. In European Conference on Computer Vision, 2018.

[12] B. Hepp, M. Nießner, and O. Hilliges. Plan3D: Viewpoint and Trajectory Optimization for Aerial Multi-View Stereo Reconstruction. ACM Transactions on Graphics (TOG), 38(1):4, 2018.

[13] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. In ICRA, 2016.

[14] J. Martinez, R. Hossain, J. Romero, and J. Little. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In International Conference on Computer Vision, 2017.

[15] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In International Conference on 3D Vision, 2017.

[16] T. Nägele, L. Meier, A. Domahidi, J. Alonso-Mora, and O. Hilliges. Real-time planning for automated multi-view drone cinematography. 2017.

[17] T. Nägele, S. Oberholzer, S. Plüss, J. Alonso-Mora, and O. Hilliges. Real-Time Environment-Independent Multi-View Human Pose Estimation with Aerial Vehicles. 2018.

[18] E. Palazzolo and C. Stachniss. Information-driven autonomous exploration for a vision-based mav. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4:59, 2017.

[19] G. Pavlakos, X. Zhou, K. Derpanis, G. Konstantinos, and K. Daniilidis. Coarse-To-Fine Volumetric Prediction for Single-Image 3D Human Pose. In Conference on Computer Vision and Pattern Recognition, 2017.

[20] G. Pavlakos, X. Zhou, K. D. G. Konstantinos, and D. Costas. Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations. In Conference on Computer Vision and Pattern Recognition, 2017.

[21] D. Pavlo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In Conference on Computer Vision and Pattern Recognition, 2019.

[22] A. Pirinen, E. Gärtner, and C. Sminchisescu. Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction. In Advances in Neural Information Processing Systems 32, pages 3907–3917. 2019.

[23] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. In Conference on Computer Vision and Pattern Recognition, 2017.

[24] S. Prokudin, P. Gehler, and S. Nowozin. Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In European Conference on Computer Vision, pages 534–551, 2018.

[25] P. Rahimian and J. K. Kearney. Optimal Camera Placement for Motion Capture Systems. IEEE Transactions on Visualization and Computer Graphics, 23(3):1209–1221, 2016.
References

[1] A. Aissaoui, A. Ouafi, P. Pudlo, C. Gillet, Z.-E. Baarir, and A. Taleb-Ahmed. Designing a Camera Placement Assistance System for Human Motion Capture Based on a Guided Genetic Algorithm. *Virtual reality*, 22(1):13–23, 2018.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Conference on Computer Vision and Pattern Recognition*, pages 1302–1310, 2017.

[3] Y. Chao, J. Yang, B. Price, S. Cohen, and J. Deng. Forecasting Human Dynamics from Static Images. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[4] X. Chen and J. Davis. Camera Placement Considering Occlusion for Robust Motion Capture. *Computer Graphics Laboratory, Stanford University, Tech. Rep.*, 2(2.2):2, 2000.

[5] W. Cheng, L. Xu, L. Han, Y. Guo, and L. Fang. ihuman3d: Intelligent human body 3d reconstruction using a single flying camera. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1733–1741. ACM, 2018.

[6] S. Choudhury, A. K. G., Ranade, and D. Dey. Learning to gather information via imitation. In *ICRA*, 2017.

[7] J. Daudelin and M. Campbell. An adaptable, probabilistic, next-best view algorithm for reconstruction of unknown 3-d objects. *IEEE Robotics and Automation Letters*, 2(3):1540–1547, 2017.

[8] A. J. Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-Time Single Camera Slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, June 2007.

[9] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent Network Models for Human Dynamics. In *International Conference on Computer Vision*, 2015.

[10] C. Gebhardt, S. Stevsic, and O. Hilliges. Optimizing for Aesthetically Pleasing Quadrotor Camera Motion. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 37(4):90:1–90:11, 2018.

[11] B. Hepp, D. Dey, S. Sinha, A. Kapoor, N. Joshi, and O. Hilliges. Learn-To-Score: Efficient 3D Scene Exploration by Predicting View Utility. In *European Conference on Computer Vision*, 2018.

[12] B. Hepp, M. Niefner, and O. Hilliges. Plan3D: Viewpoint and Trajectory Optimization for Aerial Multi-View Stereo Reconstruction. *ACM Transactions on Graphics (TOG)*, 38(1):4, 2018.

[13] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. In *ICRA*, 2016.

[14] J. Martinez, R. Hossain, I. Romero, and J. Little. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In *International Conference on Computer Vision*, 2017.

[15] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *International Conference on 3D Vision*, 2017.

[16] T. Nageli, L. Meier, A. Domahidi, J. Alonso-Mora, and O. Hilliges. Real-time planning for automated multi-view drone cinematography. 2017.
7. Supplementary Material

7.1. Supplementary Video

The supplementary video provides a short overview of our work and summarizes the methodology and results. It includes video results of our active trajectories for both the teleportation and simulated flight cases. The video is available at [https://youtu.be/Dqv7ZJQi28o](https://youtu.be/Dqv7ZJQi28o).

7.2. The Drone Flight Model

As we mention in Section 3.3 of our main document, in order to accurately predict where the drone will be positioned after passing it a goal velocity, we have formulated a drone flight model.

**Ablation Study.** We replace our drone flight model with uniform sampling around the drone. This is illustrated in Figure [1]. We evaluate the performance of our active decision making policy with the uniform sampling in Table [3]. The trajectories found using this sampling policy is shown in Figure [12]. We find that the algorithm cannot find the constant rotation policy when we remove the drone flight model and in turn, performs worse.

7.3. Results with Openpose and Liftnet

We evaluate our results on the toy example case, using the networks of [2] and [32] to find the 2D pose detections M and 3D relative pose detections L. The results are reported in Table [4]. We outperform the baselines significantly for the real image dataset MPI-INF-3DHP. For the synthetic
### Table 3. Ablation study on the importance of having a drone flight model.

We show 3D pose accuracy on simulated drone flight using noisy ground truth for estimating M and L. We show that we have a large improvement when we use our flight model to predict the future locations of the drone. Using a flight model allows us to find the same trajectories as constant rotation.

|                      | CMU-Dribble | CMU-Dance | CMU-Sitting | CMU-Dino | Total  |
|----------------------|-------------|-----------|-------------|----------|--------|
| Active with Flight Model | 0.28±0.006  | 0.15±0.007 | 0.12±0.003  | 0.18±0.01 |        |
| Active w/o Flight Model | 0.65±0.09   | 0.48±0.09  | 0.22±0.07   | 0.45±0.08 |        |
| Constant Rot.        | 0.30±0.02   | 0.15±0.01  | 0.15±0.01   | 0.20±0.02 |        |

### Figure 11. The predicted future positions of the drone (a) without using our flight model and (b) using our flight model.

### Figure 12. The trajectories drawn by our active decision making policy (a) without using our flight model and (b) using our flight model.

We outperform the baselines significantly in the real image dataset as compared to the synthetic datasets because the error of network [2] for real data is much lower than for synthetic data. We verify this by comparing the normalized 2D-pose estimation errors of a synthetic sequence and a sequence taken from the MPI-INF-3DHP dataset. We find that the normalized average error of [2] of the synthetic sequence is 0.10 with 0.08 standard deviation, whereas the normalized average error of the real image sequence is 0.06 with 0.06 standard deviation. Therefore, the unrealistically high noise of OpenPose on the synthetic data deprives strong conclusions from the first three columns of Table 4.

### Table 4. 3D pose accuracy on toy experiment, using [2][32] for estimating M and L. We outperform all predefined baseline trajectories for the real image dataset, MPI-INF-3DHP. As for the cases with synthetic input, we achieve comparable results with random, albeit with much lower standard deviation.

|                      | CMU-Walk | CMU-Dance | CMU-Run | MPI-INF-3DHP | Total  |
|----------------------|----------|-----------|---------|--------------|--------|
| Oracle (Ours)        | 0.15±0.005 | 0.25±0.009 | 0.25±0.002 | 0.21±0.008 | 0.15±0.004 |
| Random               | 0.37±0.004 | 0.24±0.005 | 0.24±0.005 | 0.42±0.005 | 0.21±0.002 |
| Constant Angle       | 0.71±0.50  | 0.76±0.37  | 0.69±0.22  | 1.26±0.53  | 0.72±0.04  |

7.4. Further Details About Simulation Environment

To test our algorithms we use the AirSim [30] drone simulator, a plug-in built for the Unreal game engine.

AirSim provides a Python API which can be used to control the drone realistically, since it uses the same flight controllers as used on actual drones. The position and orientation of the drone can be retrieved from the simulator according to the world coordinate system, which takes the drone’s starting point as the origin. The drone can be commanded to move to a with a specified velocity for a specified duration. We have added functionalities to the simulator to control a human character, get ground truth information about the character and animate it with motions from the CMU Graphics Lab Motion Capture Database [39].

For experiments requiring teleportation we use the simulator in ”ComputerVision” mode, whereas for experiments simulating flight we use ”Multirotor” mode.