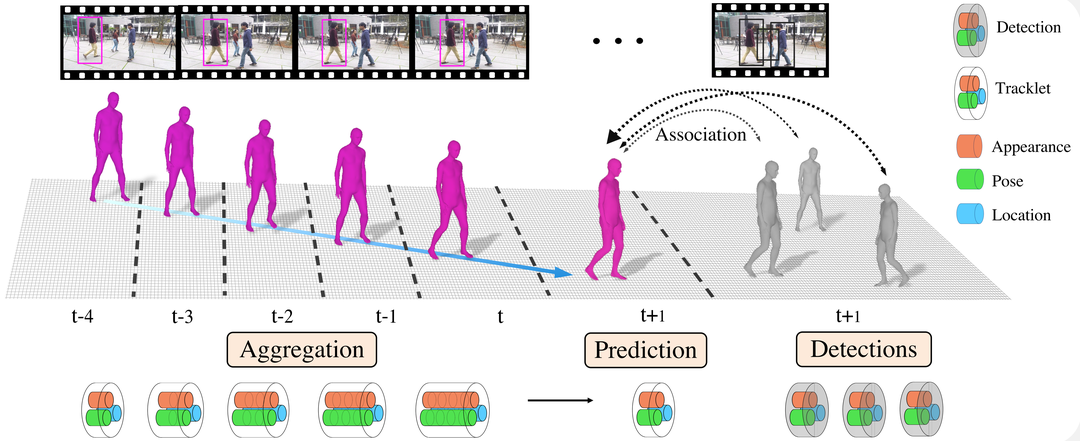
Tracking People by Predicting 3D Appearance, Location and Pose

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Jitendra Malik

UC Berkeley

Introduction

- We track people in monocular videos.
- we lift every people in a video to 3D. *Rajasegaran et al. NeurIPS 2021*
- We aggregate the states and make predictions.

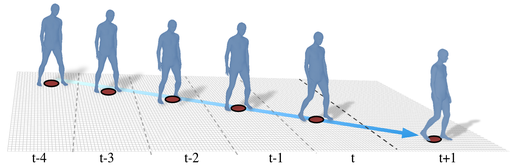


Experiments and Results

Method	Posetrack				MuPoTS				AVA	
	IDs↓	MOTA↑	IDF1↑	HOTA↑	IDs↓	MOTA↑	IDF1↑	HOTA↑	IDs↓	IDF1↑
Trackformer [29]	1263	33.7	64.0	46.7	43	24.9	62.7	53.2	716	40.9
Tracktor [4]	702	42.4	65.2	38.5	53	51.5	70.9	50.3	289	46.8
AlphaPose [8]	2220	36.9	66.9	37.6	117	37.8	67.6	41.8	939	41.9
FlowPose [39]	1047	15.4	64.2	38.0	49	21.4	67.1	43.0	452	52.9
T3DP [32]	655	55.8	73.4	50.6	38	62.1	79.1	59.2	240	61.3
PHALP	**541**	**58.9**	**76.4**	**52.9**	**22**	**66.2**	**81.4**	**59.4**	**227**	**62.7**

Table 2. **Comparison with state-of-the-art tracking methods.** We compare our method PHALP with various tracking methods in three different datasets. Our approach outperforms the other baselines across all datasets and metrics.

Location Prediction



- Human motion is linear.
- Place every human in X, Y and N.
- nearness = 1/scale.

$$\mathcal{P}_{XY}(D_j \in T_i | d_{xy} = \Delta_{xy}) \propto \frac{1}{\beta_{xy}} \exp\left(\frac{-\Delta_{xy}}{\beta_{xy}\delta_{xy}}\right)$$

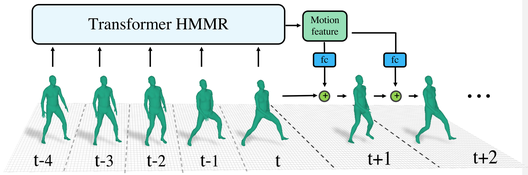
$$\mathcal{P}_{N}(D_j \in T_i | d_{n} = \Delta_{n}) \propto \frac{1}{\beta_{n}} \exp\left(\frac{-\Delta_{n}}{\beta_{n}\delta_{n}}\right)$$

Appearance Prediction

- Appearance is constant over time
- Exponential moving average to aggregate.
- Δ_α is the distance between prediction and detection.

$$\mathcal{P}_{A}(D_j \in T_i | d_{a} = \Delta_{a}) \propto \frac{1}{1 + \beta_a \Delta_a}$$

Pose Prediction

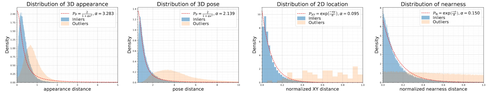


- Human Mesh and Motion Recovery to predict human poses. *Kanazawa et al. CVPR 2019*
- Pose distribution is modeled as a Cauchy distribution.

$$\mathcal{P}_{P}(D_j \in T_i | d_{p} = \Delta_{p}) \propto \frac{1}{1 + \beta_p \Delta_p}$$

Association

- Inlier, outlier distribution for appearance, pose, 2D location, and nearness.



- Total cost is the negative log likelihood of the joint distribution.

$$\mathcal{P}(D_j \in T_i | \Delta_a, \Delta_p, \Delta_{xy}, \Delta_n) \propto \mathcal{P}_A \mathcal{P}_P \mathcal{P}_{XY} \mathcal{P}_N$$

$$\begin{aligned}\Phi_C(D_j, T_i) &= -\log(\mathcal{P}(D_j \in T_i)) \\ &= -\log(\mathcal{P}_A) - \log(\mathcal{P}_P) - \log(\mathcal{P}_{XY}) - \log(\mathcal{P}_N),\end{aligned}$$

- Hungarian to solve association.