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 person in a video to 3D.
- We aggregate the states and make predictions.

Location Prediction

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

\[
P_X(Y_j \in T_i|d_{xy} = \Delta_{xy}) \propto \frac{1}{\beta_{xy}} \exp\left(-\frac{\Delta_{xy}}{\beta_{xy}}\right) \\
P_X(Y_j \in T_i|d_{xy} = \Delta_{xy}) \propto \frac{1}{\beta_{xy}} \exp\left(-\frac{\Delta_{xy}}{\beta_{xy}}\right)
\]

Appearance Prediction

- Appearance is constant over time.
- Exponential moving average to aggregate.
- \(\Delta_a\) is the distance between prediction and detection.

\[
P_a(D_j \in T_i|d_a = \Delta_a) \propto \frac{1}{1 + \beta_a \Delta_a}
\]

Pose Prediction

- Human Mesh and Motion Recovery to predict human poses.
- Pose distribution is modeled as a Cauchy distribution.

\[
P_p(D_j \in T_i|d_p = \Delta_p) \propto \frac{1}{1 + \beta_p \Delta_p}
\]

Association

- Inlier, outlier distribution for appearance, pose, 2D location, and nearness.
- Total cost is the negative log likelihood of the joint distribution.

\[
\Phi_c(D_j, T_i) = -\log(P(D_j \in T_i)) - \log(P_p(D_j \in T_i)) - \log(P_a(D_j \in T_i)) - \log(P_n(D_j \in T_i))
\]

- Hungarian to solve association.

Experiments and Results