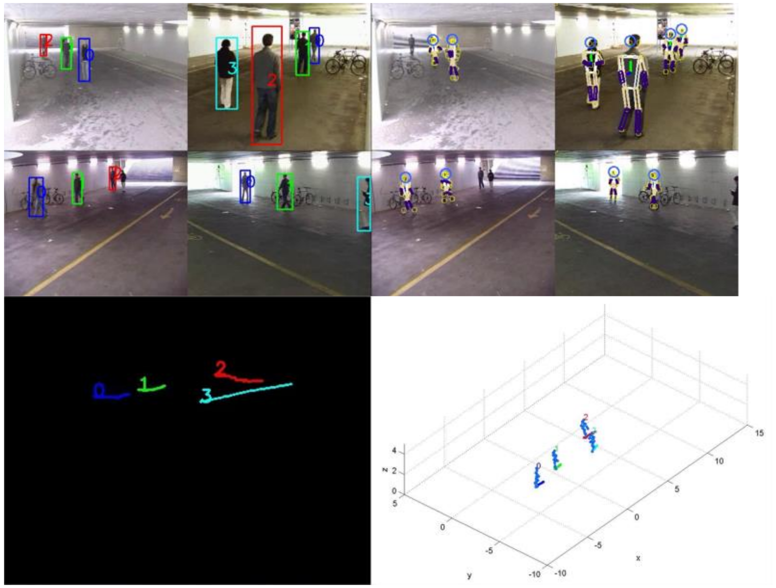
UNIVERSITY OF WASHINGTON
ELECTRICAL ENGINEERING

Joint Multi-view People Tracking and Pose Estimation for 3D Scene Reconstruction

Zheng Tang, Renshu Gu, Jenq-Neng Hwang
Information Processing Lab, Department of Electrical Engineering
University of Washington

Introduction

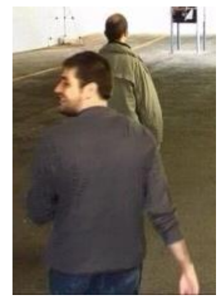


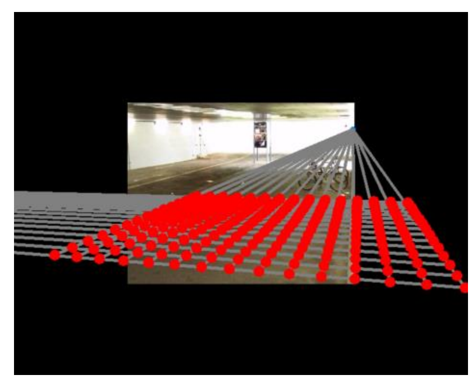
- Multi-view 3D scene reconstruction
  - 3D multiple object tracking
  - 3D human pose estimation
- 3D Multiple Object Tracking
  - Object detection + data association
- 3D Human Pose Estimation
  - Deriving 3D location of each body joint point in time

Introduction

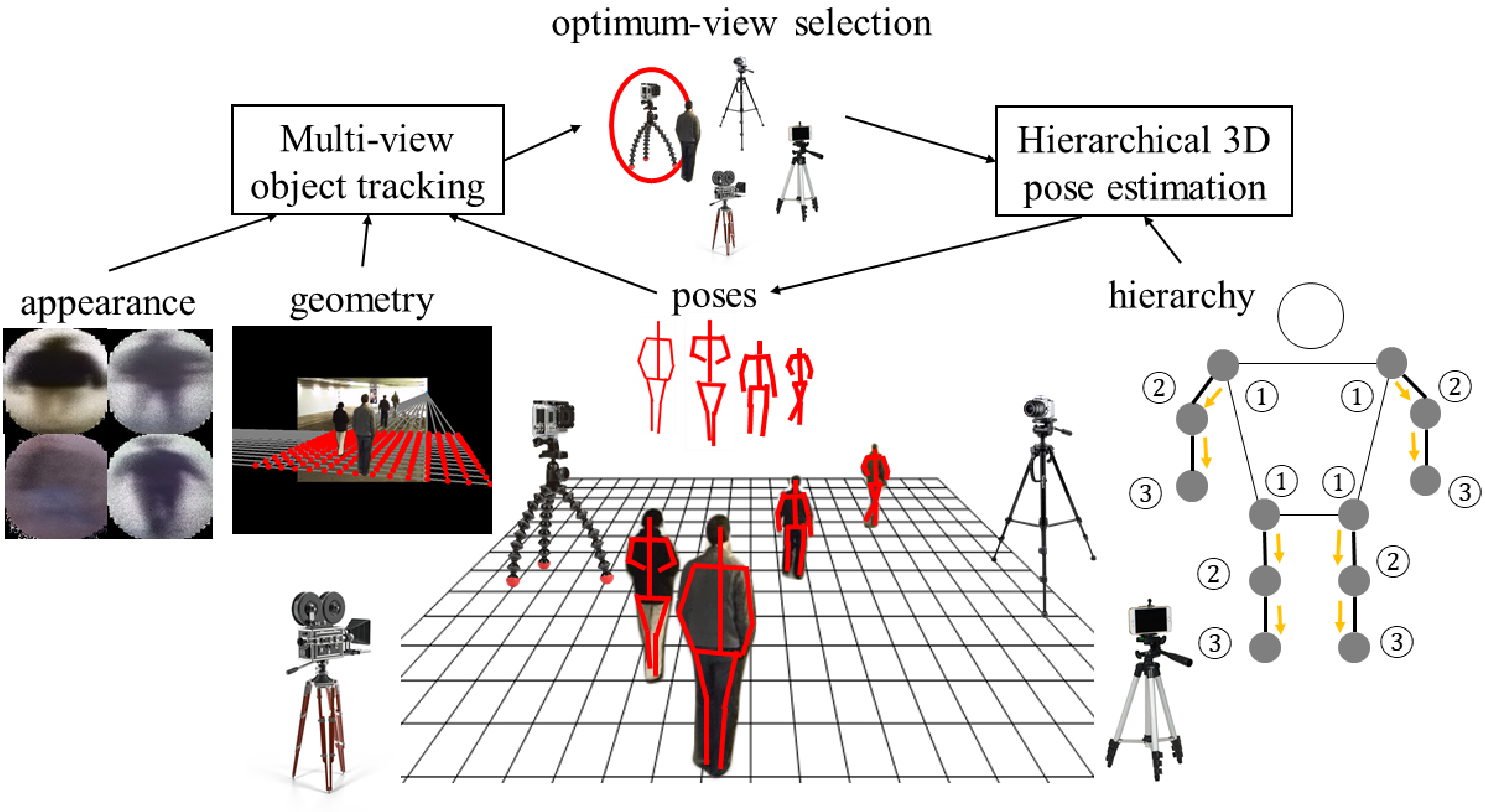
Left hand occluded by his own body

- Challenges
  - Occlusion with other objects
  - Occlusion by background
  - Self-occlusion
  - Variation in different viewing perspectives
  - Ground plane estimation / camera calibration





Overview


optimum-view selection
Multi-view object tracking
Hierarchical 3D pose estimation
appearance
geometry
poses
hierarchy
1
2
3

Multi-view Object Tracking

- Object detection by YOLO v2 [Redmon, et al. CVPR 2017]
- Tracklets formed by Kalman-filter-based tracking

$$\tau = \left\{ \left(a_j^c, g_j^c, r_j^c, t_j^c \right): j = 1,2,\dots,|\tau|, c = 1,2,\dots,C \right\}$$

- Goal: $G = \left\{ T_i \leftarrow \tau_j^c, \forall i, \forall j, \forall c \right\}$
- Solution:
  - Maximizing posterior/minimizing energy by MCMC

$$p(G|I) \propto \exp[-E(G,I)]$$

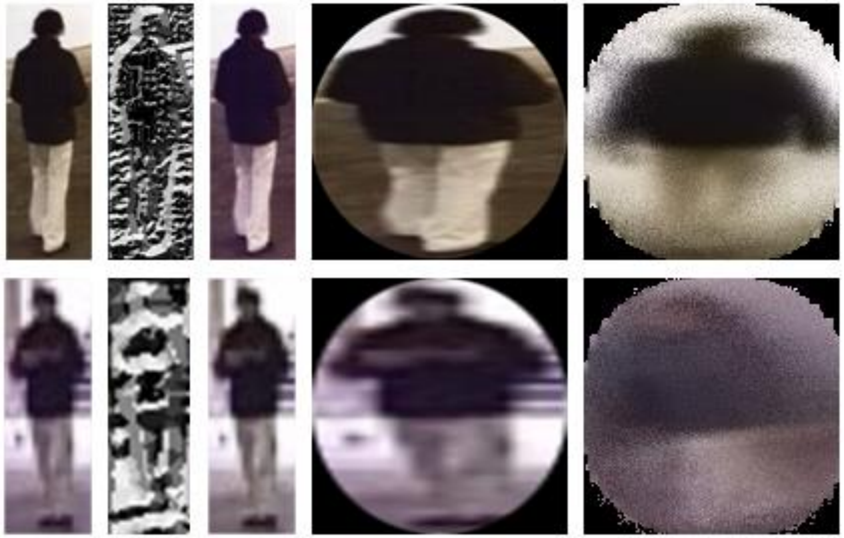
$$E(G,I) = \sum_t \left(E_t^{\text{app}} + \lambda_g E_t^{\text{geo}} + \lambda_r E_t^{\text{pos}} \right)$$

T: notation of trajectories
τ: notation of tracklets
C: number of cameras
a_j^c: adaptive appearance model
g_j^c: geometry information
r_j^c: estimated 3D human pose
t_j^c : time stamp
E_t^{app}: energy for appearance
E_t^{geo}: energy for geometry
E_t^{pos}: energy for pose
λ's: regularization parameters

Adaptive Appearance Modeling

- Concept
  - Combination of $w \times h$ pixel models
  - A history of N observed feature values at each pixel p
- Feature space: RGB (color-transformed) + LBP
- Construction
  - Normalization & color-transformation
  - Gaussian spatial weighted learning rate
    - $\alpha(p) = \exp\left[-\frac{\|p-p_c\|_2{}^2}{2(w^2+h^2)}\right]$
    - p_c – center of mass

$$a_j^c = \{a_{j,1}^c(p), a_{j,2}^c(p) \ldots, a_{j,N}^c(p)\}$$



(a) (b) (c) (d) (e)

(a) RGB images
(b) LBP images
(c) Color-transferred images
(d) Normalized bounding boxes with ellipse masks
(e) (Averaged) appearance models (color components only)

Adaptive Appearance Modeling

$$\boxed{a_j^u = \{a_{j,1}^u(p), a_{j,2}^u(p) \dots, a_{j,N}^u(p)\}}$$

- Comparison
  - Appearance model a_j^u in camera view u
  - Detected bounding box i_k^v in camera view v (color-transformed)
  - Matching/similarity score: $s_{j,k}^{u,v} = \frac{\sum_p \left[\#\left\{\left\| i_k^v(p) - a_{j,n}^c(p) \right\|_2 < \epsilon_a, \forall n < N\right\}\right]}{N \cdot w \cdot h}$
  - ϵ_a - Maximum feature distance threshold
- Energy for appearance affinity using two-way comparison
  - $E_t^{\text{app}} = \sum_i \sum_{u,v} \frac{1}{s_{j,k}^{u,v} + s_{k,j}^{v,u}}, T_i \leftarrow \tau_j^u, \tau_k^v$

T: notation of trajectories
τ: notation of tracklets

Geometry Information

- Constitution
  - $g_j^c = \left(l_j^c, d_j^c, v_j^c, b_j^c\right)$
  - l_j^c: Predicted 3D ground location in the global coordinate system
  - d_j^c: Depth to the camera c
  - v_j^c : Visibility
    - The percentage of visible area when an object is occluded by other(s)
  - b_j^c : Whether the bounding box is attached to a frame border
- Energy for geometry
  - $E_t^{\text{geo}} = \sum_i \sum_{u,v} \left[\left\| l_j^u - l_k^v \right\|_2 \cdot \frac{\min\left\{v_j^u, v_k^v\right\} \cdot b_j^u \cdot b_k^v}{\max\left\{d_j^u, d_k^v\right\}} \right], T_i \leftarrow \tau_j^u, \tau_k^v$

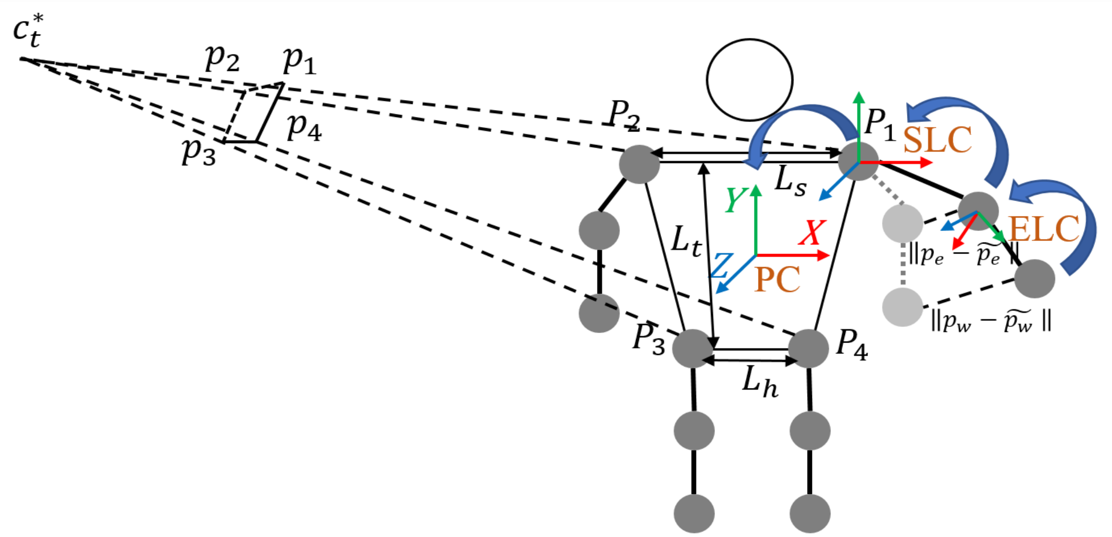
Feedback Loops

- Feedback from pose estimation to multi-view tracking
  - r_j^c - The feedback of 3D human joint points
  - Energy for pose/action attributes
    - $E_t^{\text{pos}} = \sum_i \sum_{u,v} \left[\left\| r_j^u - r_k^v \right\|_2 \cdot \frac{\min\{v_j^u, v_k^v\} \cdot b_j^u \cdot b_k^v}{\max\{d_j^u, d_k^v\}} \right], T_i \leftarrow \tau_j^u, \tau_k^v$
- Feedback from multi-view tracking to pose estimation
  - Optimum camera view selection in each frame
  - $c_t^* = \arg \max_{\forall c \leq C} \frac{v_t^c \cdot b_t^c}{d_t^c}$

C: number of cameras
d_t^c: depth
v_t^c: visibility
b_t^c: whether attached to frame border

Hierarchical 3D Pose Estimation

- Select the optimum view c_t^* for human pose estimation.
- Utilize state-of-the-art 2D pose estimation[Cao et al., CVPR 2018].
- Hierarchy
  - Torso estimation in the *person's coordinates* (PC)
  - Upper limb estimation in the *shoulder local coordinates* (SLC)
  - Lower limb estimation in the *elbow local coordinates* (ELC)



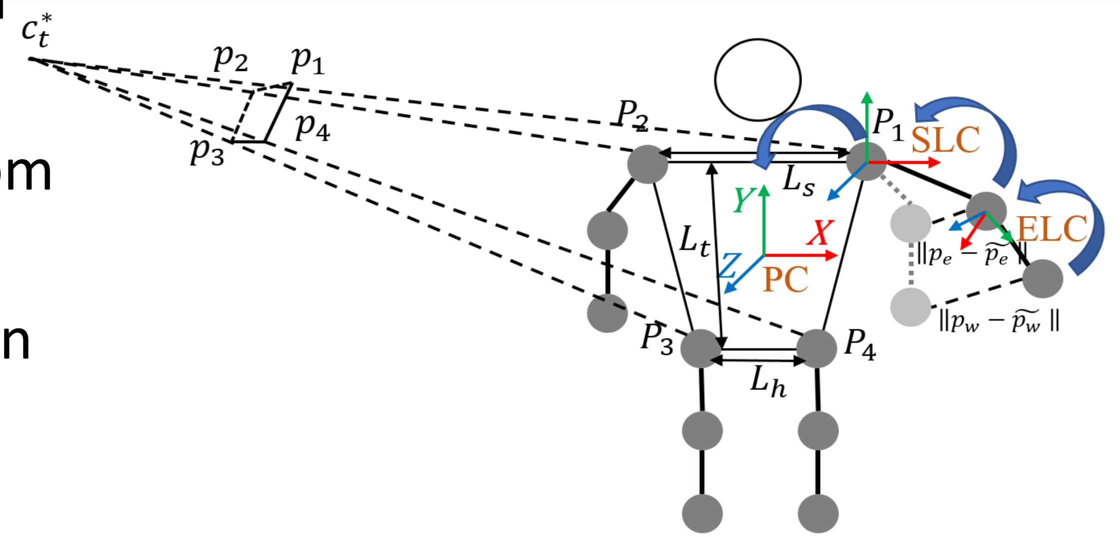
Hierarchical 3D Pose Estimation

- Torso pose estimation in the PC
  - P4P problem based on a human model prior
  - $P_1 = \left(\frac{L_s}{2}, \frac{L_t}{2}, 0\right), P_2 = \left(-\frac{L_s}{2}, \frac{L_t}{2}, 0\right), P_3 = \left(\frac{L_h}{2}, -\frac{L_t}{2}, 0\right), P_4 = \left(-\frac{L_h}{2}, -\frac{L_t}{2}, 0\right)$
- Limb estimation in the ELC and SLC
  - Wrist coordinates in ELC: $P_w^{\mathrm{ELC}} = \mathbf{R}_l^Y \mathbf{R}_l^X [0\ L_l\ 0]^T$
  - Elbow coordinates in SLC: $P_e^{\mathrm{SLC}} = \mathbf{R}_u^Z \mathbf{R}_u^Y \mathbf{R}_u^X [0\ L_u\ 0]^T$
  - Wrist coordinates in SLC: $P_w^{\mathrm{SLC}} = \mathbf{R}_u^Y \mathbf{R}_u^X \left(P_w^{\mathrm{ELC}} + [0\ L_u\ 0]^T\right)$
  - Wrist/elbow coordinates in PC
    - $P_e^{\mathrm{PC}} = P_e^{\mathrm{SLC}} + [X_s, Y_s, Z_s]^T$
    - $P_w^{\mathrm{PC}} = P_w^{\mathrm{SLC}} + [X_s, Y_s, Z_s]^T$

L_l/L_u: length of lower/upper arm
$\theta_l^X/\theta_l^Y/\theta_u^X/\theta_u^Y/\theta_u^Z$: angles to be estimated
$\mathbf{R}_l^X/\mathbf{R}_l^Y/\mathbf{R}_u^X/\mathbf{R}_u^Y/\mathbf{R}_u^Z$: rotation matrices
$P_s^{\mathrm{PC}} = [X_s, Y_s, Z_s]$: shoulder coordinates in PC

Hierarchical 3D Pose Estimation

- Limb estimation using optimization
  - Minimization of reprojection errors solved by Powell's conjugate direction method[Powell, Comput. J. 1964]
  - $f\left(\theta_u^X, \theta_u^Y, \theta_u^Z, \theta_l^X, \theta_l^Y\right) = \lambda_e \| p_e - \widetilde{p_e} \|_2 + \lambda_w \| p_w - \widetilde{p_w} \|_2$
  - p_e / p_w: back projected $P_e^{\mathrm{PC}} / P_w^{\mathrm{PC}}$
  - $\widetilde{p_e} / \widetilde{p_w}$: predictions from 2D pose estimation
  - $\lambda_e < \lambda_w$: regularization parameters



Experimental Results

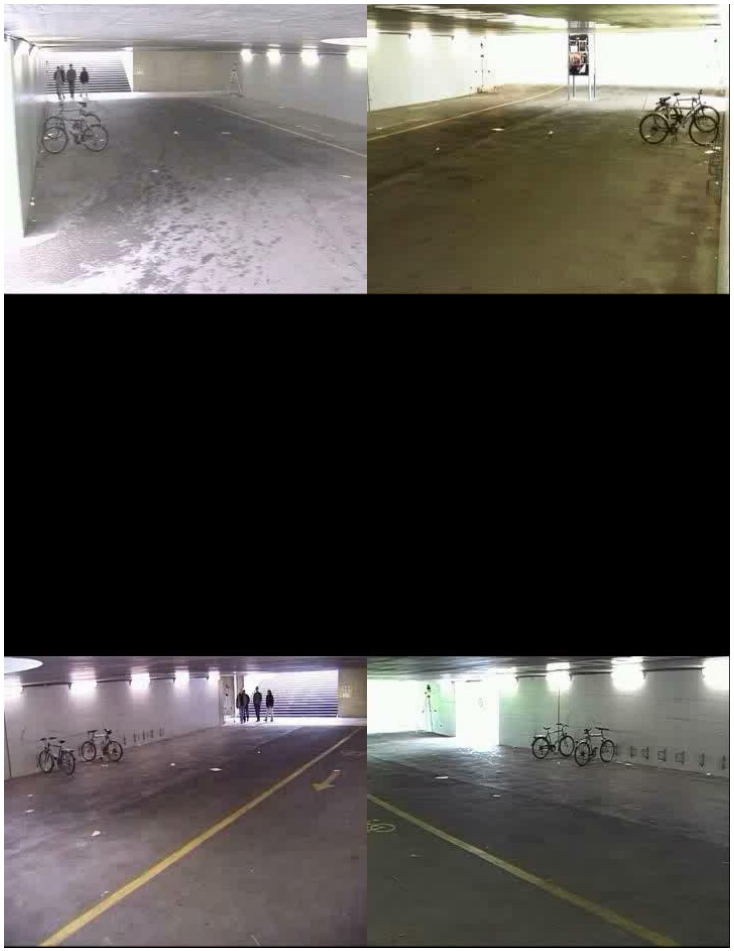
- Evaluation on EPFL benchmark[Berclaz et al., TPAMI 2011]
  - The *passageway* sequence: 4 views, 11 objects, 25 fps, 360x288
  - CLEAR metrics[Bernardin et al., EURASIP J. 2008]: *Multiple Object Detection Accuracy* (MODA), *Detection Precision* (MODP), *Tracking Accuracy* (MOTA) and *Tracking Precision* (MOTP)

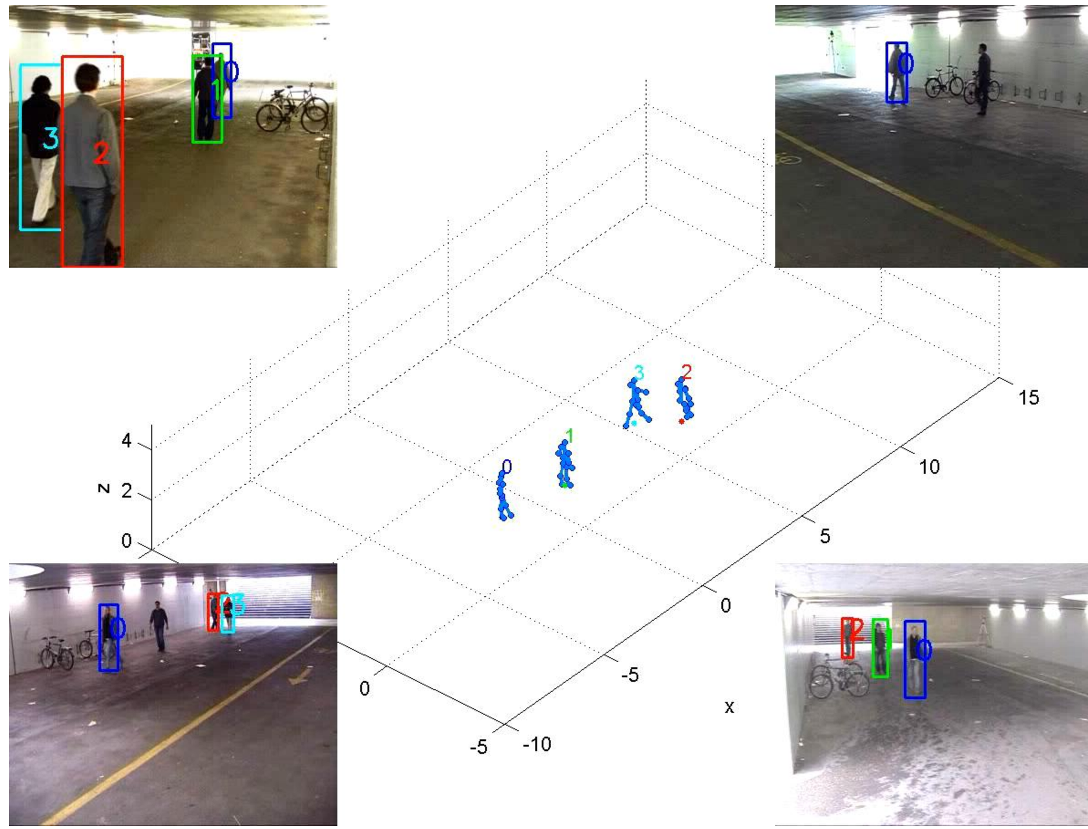
Table 1. Quantitative comparison of multi-view object tracking on the *EPFL* benchmark

Method	MODA(%)	MODP(%)	MOTA(%)	MOTP(%)
Ours	**61.04**	**73.13**	**60.26**	**72.26**
HTC [5]	43.75	67.11	43.75	67.11
KSP [2]	40.46	58.88	40.46	57.24
POM [3]	32.57	62.50	32.57	60.86

Bold entries indicate the best results in the corresponding columns.

Experimental Results





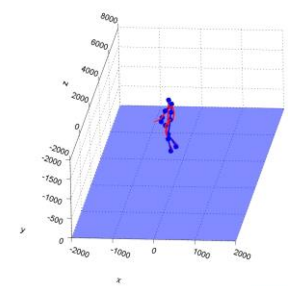
Experimental Results

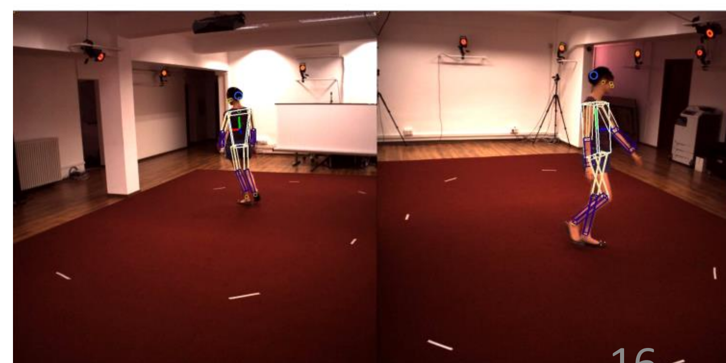
- Evaluation on Human3.6M benchmark [Ionescu et al., TPAMI 2014]
  - The *walking* sequence: 4 views, 1 object, 50 fps, 1000x1002
  - Metrics: Average 3D distance between the ground truths and the estimated joint points
  - Conclusion: The effectiveness of optimum-view selection is verified

Table 2. Quantitative comparison of 3D pose estimation on the *Human3.6M* benchmark (unit: mm)

Multi-view	Camera #0	Camera #1	Camera #2	Camera #3
99.7	132.5	115.1	113.2	137.1

Experimental Results





Conclusion

- 3D scene reconstruction combining multi-view object tracking and 3D human pose estimation
- Multi-view object tracking using appearance, geometry and pose/action attributes
- 3D human pose estimation using hierarchical optimization
- Feedback loops between tracking and pose estimation
- Successful evaluation on two benchmark datasets