# JOINT MULTI-VIEW PEOPLE TRACKING AND POSE ESTIMATION FOR 3D SCENE RECONSTRUCTION

*Zheng Tang, Renshu Gu, Jenq-Neng Hwang*

Department of Electrical Engineering, University of Washington (UW)
{zhtang, renshugu, hwang}@uw.edu

## ABSTRACT

The goal of data analytics in surveillance videos is to fully understand and reconstruct the 3D scene, i.e., to recover the trajectory and action of each object. In a surveillance system with camera arrays of overlapping views, we propose a novel video scene reconstruction framework to collaboratively track multiple human objects and estimate their 3D poses. First, tracklets are extracted from each single view following the tracking-by-detection paradigm. We propose an effective integration of visual and semantic object attributes, i.e., appearance models, geometry information and poses/actions, to associate tracklets across different views. Based on the optimum viewing perspectives derived from tracking, a hierarchical estimation of human poses is introduced to generate the 3D skeleton of each object. The estimated body joint points are fed back to the tracking stage to enhance tracklet association. Experiments on benchmarks of multi-view tracking and 3D pose estimation validate the effectiveness of the proposed method.

***Index Terms***— 3D scene reconstruction, multi-view tracking, 3D pose estimation, multiple object tracking, data association

## 1. INTRODUCTION

The growing demand of user experience with video streaming has brought about a rapid growth in big visual data analytics. The ultimate purpose of major applications in this research field is to fully understand and reconstruct the video scene in a 3D space. This not only involves accurate identification of multiple objects and the recovery of their trajectories, but also requires precise estimation of their postures.

Recently, cross-view tracking of multiple people in a surveillance camera array has attracted lots of attentions in the literature [1]. Researchers exploit multiple cues in both 2D and 3D, e.g., ground plane occupancy [2]-[4], motion coherence, appearance affinity [5], temporal consistency [6], postures and actions [7], etc., to locate multiple targets in a 3D scene map. Nonetheless, there remain many challenges that have not been fully resolved. First, in crowded scenes where people frequently occlude and intersect with one another, the number of identity switches can increase rapidly. Moreover, the same object may experience large appearance variation across different viewpoints. Last but not least, the common inaccuracy of ground plane estimation causes mistakes in geo-localization, especially for objects that are far away in a video scene.

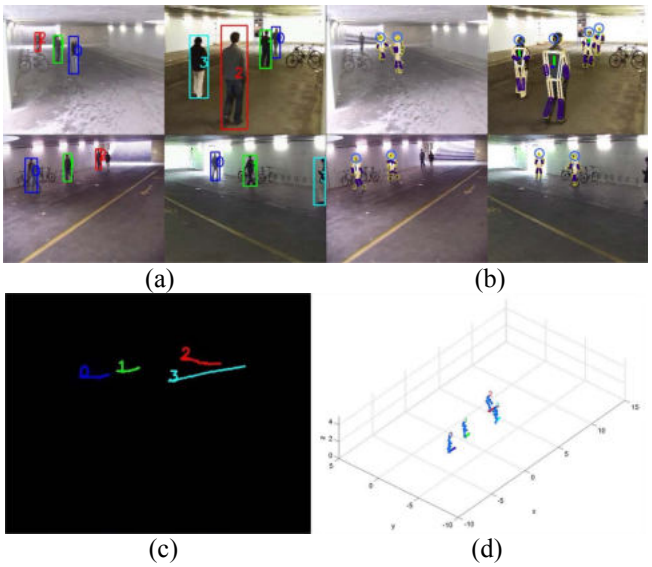



**Fig. 1.** (a) Multi-view people tracking in 2D. (b) Multi-view pose estimation. (c) top view of 3D trajectories. (d) 3D scene reconstruction.

On the other hand, the estimation of human poses is another key component to multi-view scene reconstruction. Multi-view 3D pose estimation provides informative and view-invariant features for many useful applications such as action recognition. However, full recovery of 3D human poses for multiple objects also remains unsolved in dynamic and cluttered environments. The major challenge is the under-constrained nature of the problem due to loss of depth information and frequent (self-)occlusion.

In this paper, we propose to jointly collaborate multi-view multi-object tracking and 3D human pose estimation for scene reconstruction. For initialization, we follow the tracking-by-detection paradigm to generate tracklets, which are a series of human bounding boxes grouped by spatio-temporal coherency and perceptual similarity. Then we formulate the data association problem as energy

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

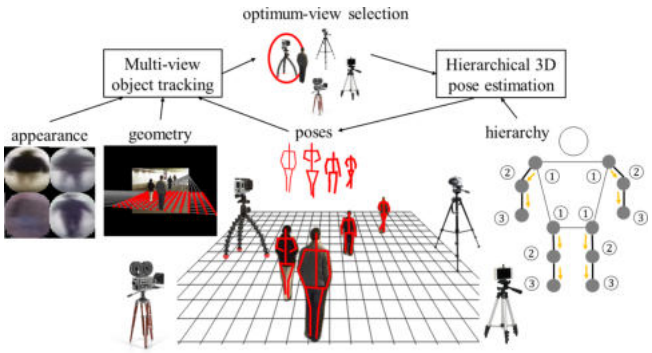



**Fig. 2.** An illustration of the multi-view scene reconstruction.

minimization based on a set of visual and semantic attributes, including a pixel-based adaptive model for two-way appearance comparison and a geometry proximity measurement based on weighting of depth and visibility. We also introduce an explicit action descriptor using feedback from the pose estimation stage. The geometry information in multi-view tracking is used to choose the optimum viewpoint for hierarchical 3D pose estimation, in which the limb pose estimation is formulated as minimization of reprojection errors at elbows and wrists. The proposed method is evaluated on the benchmark of multi-view human tracking in comparison with the state-of-the-art. We also run experiments on the benchmark of multi-view human pose estimation to validate the improvement by feedback from the tracking stage.

The main contribution of this work is two-fold. Firstly, novel representations of visual and semantic attributes are adopted in multi-view data association, which is formulated as an energy minimization problem. Secondly, we propose a hierarchical model for pose estimation, in which limb pose estimation is solved by minimization of reprojection errors.

The rest of this paper is organized as follows. We review related works in Section 2. The framework of multi-view 3D object tracking and hierarchical pose estimation is detailed in Section 3. The experimental results are presented in Section 4 and we conclude the paper in Section 5.

## 2. RELATED WORKS

This work is closely related to the research streams of multi-view object tracking and human pose estimation.

**Multi-view object tracking** is often formulated as data association across cameras. Berclaz et al. [2] and Fleuret et al. [3] follow a tracking-by-segmentation strategy to detect candidate targets. They respectively develop their data association approaches based on the *k-shortest paths algorithm* and the *hidden Markov process*. In [5], Xu et al. use tracking by detection and exploit multiple cues in their hierarchical composition model. Their appearance coherence is measured by *deep convolution neural network* (DCNN) features, while the motion information is encoded in a continuous function. In [6], Liu uses raw pixel template in appearance modeling and combine it with 3D localization, spline fitting and temporal consistency in the objective. Both appearance models in [5] and [6] cannot adaptively "memorize" past feature values. Furthermore, Xu et al. [7] first introduce pose/action attributes in cross-view association. However, their human poses are trained from DCNN features for categorization without pose estimation, which may cause errors in transitions of actions.

**3D human pose estimation** has been extensively studied in the last decade. Early approaches [8], [9] are based on appearance models, e.g., silhouettes, and stochastic search with kinematic constraints for tracking of joint points. However, silhouette extraction becomes unreliable for complex background and moving cameras. Amin et al. [10] introduce unconstrained 3D pose estimation from multiple camera views in a complex environment, but the algorithm is only suitable for the upper human body. More recently, Li et al. [11] train their single-camera pose estimator using 3D motion capture data. Some other monocular methods [12], [13] use advanced 2D pose detector as an intermediate step. The proposed hierarchical human pose estimator also exploits the state-of-the-art 2D pose detector [14] in the recovery of 3D poses. Different from previous approaches, results from multiple cameras are dynamically combined for optimum scene reconstruction based on feedback from tracking.

## 3. METHODOLOGY

Our proposed framework for multi-view scene reconstruction consists of two main steps (see Fig. 2). First, we track each target by data association across different views using multiple cues, including feedback from pose estimation. Second, his/her 3D body skeleton is computed in a hierarchical formulation using geo-localization information from multi-view tracking.

### 3.1. Multi-view object tracking by data association

In each single view, we first make use of the state-of-the-art object detector [15] to obtain the detected bounding boxes at each frame. Then we employ a Kalman-filter-based approach [16] to associate them into tracklets. Specifically, each trajectory is fragmented either when it (1) exits from a frame border, (2) is occluded, or (3) has a Kalman prediction of 3D location that is 1 meter away from the closest location. All the cameras are self-calibrated based on a set of 2D tracklets [17] and converted to a global coordinate system according to some shared reference point(s).

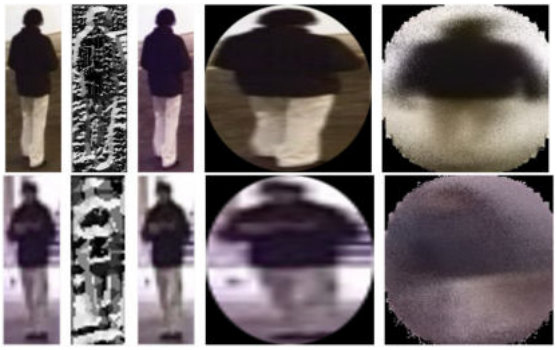
Our objective is to recover the trajectories $T$ of all people within the 3D scene, that is,

$$T = \{T_i : i = 1,2,\dots,|T|\}. \quad (1)$$

Tracklets $\tau$ are the basic units in multi-view tracking, consisting of appearance, geometry and posture information, in a time period across multiple cameras.

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

where $a_j^c$, $g_j^c$, $r_j^c$ and $t_j^c$ respectively denote the appearance feature, geometry information, estimated 3D skeleton and the



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

**Fig. 3.** Adaptive appearance models and their comparison with detected objects. The two rows give the same object identity in two different views. (a) RGB images. (b) LBP images. (c) Color-transferred images. (d) Normalized bounding boxes with ellipse masks. (e) (Averaged) appearance models (color components only).

time stamp, $c$ indexes each camera and $C$ is the total number of cameras. All the false positives of trajectories are collected at $\tau_0^c$. We aim to solve the following representation:

$$G = \{T_i \leftarrow \tau_j^c, \forall i, \forall j, \forall c\}, \quad (3)$$

which can be formulated as searching for the optimal solution by maximizing a posterior

$$p(G|I) \propto \exp[-E(G,I)], \quad (4)$$

where $I$ denotes the input image sequences and $E(G,I)$ is the total energy function over three semantic attributes (appearance, geometry and pose):

$$E(G,I) = \sum_t \left(E_t^{\text{app}} + \lambda_g E_t^{\text{geo}} + \lambda_r E_t^{\text{pos}}\right), \quad (5)$$

where $\lambda_g$ and $\lambda_r$ are regularization parameters. This energy minimization problem in (5) can be effectively solved by the reversible jump *Markov Chain Monte Carlo* (MCMC) method [18]. The gaps between associated tracklets are interpolated linearly.

### *3.1.1. Appearance attributes*

The term $E_t^{\text{app}}$ is defined to describe appearance affinity of detected bounding boxes. We propose to model the appearance of each target based on an adaptive scheme. The term $a_j^c$ is defined by a combination of $w \times h$ pixel models, which each "memorizes" a history of $N$ observed feature values at each corresponding pixel location $p$:

$$a_j^c = \{a_{j,1}^c(p), a_{j,2}^c(p) \ldots, a_{j,N}^c(p)\}. \quad (6)$$

The procedure of model construction and update is described in Fig. 3. The object region within each detected bounding box is normalized to $w \times h$ pixels masked with a maximum ellipse. A Gaussian spatial weighting scheme is introduced to dynamically control the learning rates $\alpha(p)$.

$$\alpha(p) = \exp\left[-\frac{\|p-p_c\|_2^{\,2}}{2(w^2+h^2)}\right], \quad (7)$$

where $p_c$ is the center of mass of the object region. Therefore, we can reduce the influence of background area that is usually far from the center. In each frame, if there are less than $N$ feature vectors at a pixel $p$ in $a_j^c$, the observed feature vector at $p$ is added to $a_j^c$ by a probability of $\alpha(p)$. Otherwise, a random feature vector $a_{j,n}^c(p)$ is swapped by the observed feature vector with a probability of $\alpha(p)$.

Let $u$ and $v$ denote two different views. To compare a detected box $i_k^v$ at the beginning of a tracklet in one view with an appearance model constructed in another view $a_j^u$, we adopt the color transfer method used in inter-camera tracking [19], [20] to compensate for the change of illumination and color response across different cameras. The matching score of appearance similarity is computed as

$$s_{j,k}^{u,v} = \frac{\sum_p \left[\#\left\{\left\|i_k^v(p) - a_{j,n}^c(p)\right\|_2 < \epsilon_a, \forall n < N\right\}\right]}{N \cdot w \cdot h}, \quad (8)$$

where $\epsilon_a$ is the maximum feature distance threshold. (8) measures the sum of matched samples within the object region weighted by the total number of samples. Hence, we can define the objective energy for appearance affinity as

$$E_t^{\text{app}} = \sum_i \sum_{u,v} \frac{1}{s_{j,k}^{u,v} + s_{k,j}^{v,u}}, T_i \leftarrow \tau_j^u, \tau_k^v, \quad (9)$$

in which we utilize two-way comparison to enhance the robustness of the appearance descriptor.

In experiments, the RGB color values and *local binary pattern* (LBP) values are adopted as features for appearance modeling. The absolute color distance threshold and LBP distance threshold are both set to 30. The dimension of each appearance model is 128x128x16.

### *3.1.2. Geometry attributes*

The term $E_t^{\text{geo}}$ encourages to minimize the distance of each pair of object locations assigned to the same object identity.

The geometry information of $\tau_j^c$ covers four aspects.

$$g_j^c = \left(l_j^c, d_j^c, v_j^c, b_j^c\right), \quad (10)$$

where $l_j^c \in \mathbb{R}^2$ is the predicted 3D ground location in the global coordinate system, $d_j^c$ is the depth to the camera, $v_j^c$ is the visibility defined as the percentage of visible area when an object is occluded by other(s) [21], and $b_j^c$ is an indicator of whether the bounding box is attached to a frame border. $b_j^c$ is set to 1 when the shortest distance of an edge to a frame border is larger than 10 pixels and 0.01 otherwise. The objective energy for geometry can be defined accordingly.

$$E_t^{\text{geo}} = \sum_i \sum_{u,v} \left[\left\|l_j^u - l_k^v\right\|_2 \cdot \frac{\min\{v_j^u, v_k^v\} \cdot b_j^u \cdot b_k^v}{\max\{d_j^u, d_k^v\}}\right], T_i \leftarrow \tau_j^u, \tau_k^v, \quad (11)$$

where $\left\|l_j^u - l_k^v\right\|_2$ indicates the Euclidean distance between $l_j^u$ and $l_k^v$. The 3D distance is divided by the maximum depth between two detected objects, because the precision of 3D localization decreases as an object moves far away from the camera. Moreover, since the estimation of foot points is prone to error when a bounding box is occluded or attached to a frame border, the objective is multiplied by the minimum visibility and the indicators of attachment to frame borders.

### *3.1.3. Pose attributes*

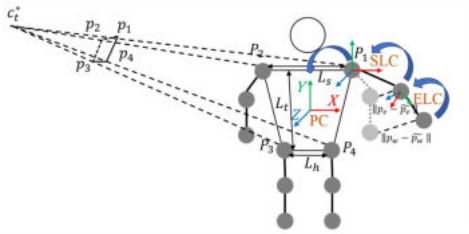



**Fig. 4.** Perspective 4 Points for torso estimation and the demonstration of limb pose estimation.

The feedback of 3D human joint points, noted $r_j^c$, from pose estimation is applied to the objective energy as pose/action attributes. Different from the previous work [7], our pose descriptor is a set of joint points encoding the 3D actions explicitly. This can help avoid confusion in pose transitions and reduce complexity with temporal information. The pose objective is weighted by some geometry aspects similar to (11), because the estimation of human skeleton also relies on high resolution and high visibility. More specifically, $E_t^{\text{pos}}$ is defined as follows.

$$E_t^{\text{pos}} = \sum_i \sum_{u,v} \left[ \left\| r_j^u - r_k^v \right\|_2 \cdot \frac{\min\{v_j^u, v_k^v\} \cdot b_j^u \cdot b_k^v}{\max\{a_j^u, a_k^v\}} \right], T_i \leftarrow \tau_j^u, \tau_k^v, \quad (12)$$

where $\left\| r_j^u - r_k^v \right\|_2$ measures the Euclidean distance between the set of joint points shared in both views.

### 3.2. Hierarchical 3D pose estimation

The hierarchy of 3D pose estimation is ordered by torso estimation, upper limb estimation and lower limb estimation.

From the geometry information in multi-view tracking, we can define the optimum view of each $T_i$ at time $t$ as

$$c_t^* = \arg \max_{\forall c \le C} \frac{v_t^c \cdot b_t^c}{d_t^c}, \quad (13)$$

which is chosen for single-view 3D pose estimation. For each camera, the camera intrinsic matrix is assumed to be self-calibrated in advance [17]. We aim to find the human joint points in the *person's coordinates* (PC), where the origin is located at the center of torso, X-axis points to the right, Y-axis points upwards and Z-axis points towards the camera (see Fig. 4). Then PC are converted to the world coordinate system, in which estimated joint points can be augmented onto the 3D trajectories for scene reconstruction.

We take advantage of the recent advance in 2D human pose estimation using a DCNN [14], which gives 2D joints prediction at every frame. The 2D predictions are utilized as input to our 3D estimation algorithm.

#### *3.2.1. Torso pose estimation*

The estimation of each person's torso pose with respect to the camera is formulated as a *perspective N point* (PNP) problem. We employ a human model prior that contains all the bone lengths with reasonable ratios between every pair of bones. The section of human torso is modeled as a trapezoid with upper base of $L_s$, height of $L_t$ and lower base of $L_h$ (see Fig. 4). Thus, the four torso joint points in *person's coordinates* (PC) can be defined as

$$P_1 = \left(\frac{L_s}{2}, \frac{L_t}{2}, 0\right), P_2 = \left(-\frac{L_s}{2}, \frac{L_t}{2}, 0\right),$$
$$P_3 = \left(\frac{L_h}{2}, -\frac{L_t}{2}, 0\right), P_4 = \left(-\frac{L_h}{2}, -\frac{L_t}{2}, 0\right). \quad (14)$$

The 2D predictions of shoulders and hips from DCNN are regarded as the four corner points of the trapezoid, i.e. 2D projection of $P_i, i = 1,2,3,4$. This P4P problem can be solved based on the projection relationship from 3D to 2D.

#### *3.2.2. Limb pose estimation*

Once the torso poses are estimated, we can move on to the next level in hierarchical 3D human pose estimation. For each limb, we aim to find the Euler angles at upper (shoulder/hip) and lower (elbow/knee) parts. The problem is formulated as minimization of reprojection errors via optimization. Inspired by [22], we define the *elbow local coordinates* (ELC), where the origin is at an elbow joint point and the axes are in accordance with PC. The length of the lower arm is denoted as $L_l$ and the angles to be estimated are $\theta_l^X$ and $\theta_l^Y$. Thus, the wrist coordinates $P_w$ in ELC can be calculated by

$$P_w^{\text{ELC}} = \mathbf{R}_l^Y \mathbf{R}_l^X [0\ L_l\ 0]^T, \quad (15)$$

where $\mathbf{R}_l^X$ and $\mathbf{R}_l^Y$ are the rotation matrices with $\theta_l^X$ and $\theta_l^Y$ respectively. Likewise, we define the *shoulder local coordinates* (SLC), where the origin locates at a shoulder point. We denote upper arm length as $L_u$ and the angles to be estimated as $\theta_u^X$, $\theta_u^Y$ and $\theta_u^Z$. The elbow coordinates in SLC are given by

$$P_e^{\text{SLC}} = \mathbf{R}_u^Z \mathbf{R}_u^Y \mathbf{R}_u^X [0\ L_u\ 0]^T. \quad (16)$$

The wrist is constrained by the elbow as

$$P_w^{\text{SLC}} = \mathbf{R}_u^Y \mathbf{R}_u^X (P_w^{\text{ELC}} + [0\ L_u\ 0]^T). \quad (17)$$

The shoulder coordinates in PC, namely $P_1, P_2$ in (14), are noted as $P_s^{\text{PC}} = [X_s, Y_s, Z_s]$, which can be used to transform the elbow and wrist points to PC.

$$P_e^{\text{PC}} = P_e^{\text{SLC}} + [X_s, Y_s, Z_s]^T,$$
$$P_w^{\text{PC}} = P_w^{\text{SLC}} + [X_s, Y_s, Z_s]^T. \quad (18)$$

3D coordinates $P_e^{\text{PC}}$ and $P_w^{\text{PC}}$ are then projected to 2D image using the projection matrices solved in torso estimation, noted as $p_e$ and $p_w$ respectively, for the computation of reprojection errors. We denote the input 2D predictions at elbow as $\widetilde{p_e}$ and at wrist as $\widetilde{p_w}$. Therefore, the objective function of this optimization problem can be formulated as

$$f(\theta_u^X, \theta_u^Y, \theta_u^Z, \theta_l^X, \theta_l^Y) = \lambda_e \| p_e - \widetilde{p_e} \|_2 + \lambda_w \| p_w - \widetilde{p_w} \|_2, \quad (18)$$

where the weights are constrained by $\lambda_e < \lambda_w$, because the computation of a wrist is affected by the elbow. We make use of the Powell's conjugate direction method [23] to solve this optimization problem efficiently.

## 4. EXPERIMENTAL RESULTS

Our proposed method is evaluated and demonstrated on the *EPFL* benchmark [2] and the *Human3.6M* benchmark [24].

### 4.1. Evaluation on EPFL benchmark

**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.

We adopt the *passageway* sequence in our experiments, which is known for its challenging scenario with poor lighting and image quality. People can become very small on the far end and some of them are captured in only one or two cameras. The sequence consists of 4 different views and films 11 pedestrians walking or bicycling. Each view is shot at 25 fps and in a relatively low resolution 360x288.

The quantitative comparison of the proposed method with several state-of-the-art algorithms in multi-view object tracking is presented in Table 1. The widely used CLEAR metrics [25] are adopted, including *Multiple Object Detection Accuracy* (MODA), *Detection Precision* (MODP), *Tracking Accuracy* (MOTA) and *Tracking Precision* (MOTP). MODA and MOTA measure three sources of errors in detection and tracking respectively: false positives, false negatives and identity switches. MODP and MOTP are used to measure misalignment between annotated and predicted locations.

The proposed algorithm achieves the top performance on all metrics in this challenging sequence. Our promising performance in tracking is mainly due to the effective formulation of multi-view object tracking by integrating robust visual and semantic attributes including appearance, geometry and human poses. HTC [5] loses to us by margin, as they only consider appearance and motion patterns in their hierarchical feature model. Moreover, their feature descriptor for appearance modeling is only extracted in several frames, however, the proposed adaptive appearance model can "memorize" a rather long history of past feature values. Both the works [2] and [3] rely on a less robust appearance representation and suffer from an object localization strategy of poor accuracy. The relatively higher MODA and MODP gained by our method and HTC [5] confirm that the tracking-by-detection-based approaches are superior to tracking by segmentation in terms of localization of object observations. The DCNN framework for object detection [8] chosen by us is also more advanced than the previous architecture [26].

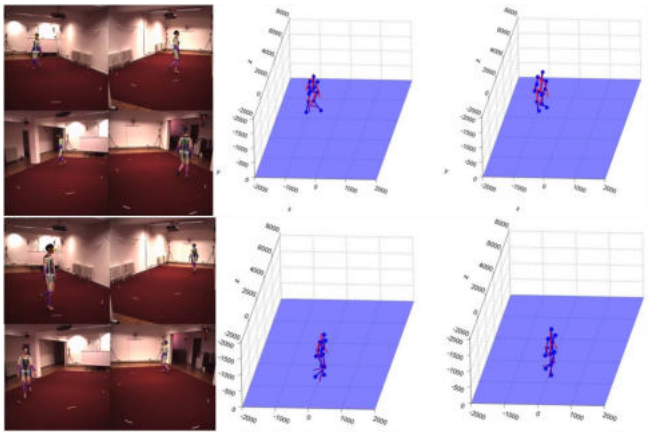
A qualitative demonstration of the proposed framework for multi-view scene reconstruction can be seen in Fig. 1. The complete video demo is made available at the following link: http://allison.ee.washington.edu/thomas/mvsr/.

### 4.2. Evaluation on Human3.6M benchmark

As there is no ground truth of 3D joint points provided in the *EPFL* benchmark, we conduct more experiments on a *Human3.6M* dataset to evaluate our improvement in 3D pose estimation. The *walking* sequence, which includes 4 high-resolution views (1000x1002) of video data at 50 fps, is chosen for experiments. There is only one object walking around the room for about 1 minute.

**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 |



(a) (b) (c)

**Fig. 5.** Qualitative comparison of 3D pose estimation on the *Human3.6M* benchmark. (a) 2D poses estimated from each view. (b) Failures of 3D pose estimation in single views. (c) Estimated 3D poses from the optimum views.

We compare the proposed strategy that uses feedback from multi-view tracking to choose the optimum camera views for pose estimation with the results directly obtained from each single view. The quantitative comparison is shown in Table 2. The metric for evaluation is the average 3D distance between the ground truths and the estimated joint points. Some qualitative results and demonstration are presented in Fig. 5. The complete video demo is also available at http://allison.ee.washington.edu/thomas/mvsr/.

Our proposed scheme based on optimum-view selection from multi-view tracking achieves the minimum error, which validates the effectiveness of optimum viewpoint estimation by (13). Since the selected camera views contain object locations with small depths and not attached to any frame border, they help generate the best performance in both pose estimation and 3D scene reconstruction.

## 5. CONCLUSION

This paper presents a multi-view scene reconstruction framework jointly combining the efforts of multi-view multi-object tracking and 3D pose estimation. Multi-view people tracking is leveraged with rich visual and semantic attributes, including adaptive appearance modeling, spatially-weighted geometry measurement, and the feedback of 3D joint points

from the pose estimation stage. The data association across different views is formulated as an energy minimization problem that is solved by an MCMC-based approach. Furthermore, we introduce a hierarchical model for 3D pose estimation, which applies a novel formulation of minimization of reprojection errors to the computation of limb angles. The estimation of 3D poses is benefited from the optimum views selected from multi-view tracking. Experiments on two public benchmarks both demonstrate the efficacy of the proposed method. In the future, we will extend our method to non-human objects and camera arrays without overlapping view [27] for broader application.

## 6. REFERENCES


[1] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao and T.-K. Kim, "Multiple Object Tracking: A Literature Review," *arXiv: 1409.7618 [cs]*, Sep. 2014, arXiv: 1409.7618. [Online]. Available: http://arxiv.org/abs/1409.7618.

[2] J. Berclaz, F. Fleuret, E. Turetken and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 33, no. 9, pp. 1806-1819, 2011.

[3] F. Fleuret, J. Berclaz, R. Lengagne and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 30, no. 2, pp. 267-282, 2008.

[4] Y.-S. Lin, K.-H. Lo, H.-T. Chen and J.-H. Chuang, "Vanishing point-based image transforms for enhancement of probabilistic occupancy map-based people localization," *IEEE Trans. Image Processing*, vol. 23, no. 12, pp. 5586-5598, 2014.

[5] Y. Xu, X. Liu, Y. Liu,and S. C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pp. 4256-4265, 2016.

[6] X. Liu, "Multi-View 3D Human Tracking in Crowded Scenes," in *Proc. Conf. Assoc. Advancement Artificial Intelligence*, pp. 3553-3559, 2016

[7] Y. Xu, X. Liu, L. Qin and S. C. Zhu, "Cross-View People Tracking by Scene-Centered Spatio-Temporal Parsing," in *Proc. Conf. Assoc. Advancement Artificial Intelligence*, pp. 4299-4305, 2017.

[8] H. Sidenbladh, M. J. Black and D. J. Fleet, "Stochastic tracking of 3d human figures using 2d image motion," in *Proc. European Conf. Comput. Vis.*, pp. 702-718, 2000.

[9] J. Deutscher, A. Blake and I. D. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pp. 126-133, 2000.

[10] S. Amin, M. Andriluka, M. Rohrbach and B. Schiele, "Multi-view Pictorial Structures for 3D Human Pose Estimation," in *Proc. British Machine Vis. Conf.*, 2013.

[11] S. Li, W. Zhang and A. B. Chan, "Maximum-margin structured learning with deep networks for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2848-2856, 2015.

[12] C. Wang, Y. Wang, Z. Lin, A. L. Yuille and W. Gao, "Robust estimation of 3D human poses from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pp. 2361-2368, 2014.

[13] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. European Conf. Comput. Vis.*, pp. 561-578, 2016.

[14] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017.

[15] J. Redmon, and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017.

[16] Z. Tang, J.-N. Hwang, Y.-S. Lin and J.-H. Chuang, "Multiple-kernel adaptive segmentation and tracking (MAST) for robust object tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 1115-1119, 2016.

[17] Z. Tang, Y.-S. Lin, K.-H. Lee, J.-N. Hwang, J.-H. Chuang and Z. Fang, "Camera self-calibration from tracking of moving persons," in *Proc. IEEE Int. Conf. Pattern Recogn.*, pp. 260-265, 2016.

[18] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711-732, 1995.

[19] Y.-G. Lee, Z. Tang, J.-N. Hwang and Z. Fang, "Inter-camera tracking based on fully unsupervised online learning," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2017.

[20] Y.-G. Lee, Z. Tang, J.-N. Hwang and Z. Fang, "Online-learning-based human tracking across non-overlapping cameras," *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[21] Z. Tang, G. Wang, T. Liu, Y.-G. Lee, A. Jahn, X. Liu, X. He and J.-N. Hwang, "Multiple-kernel based vehicle tracking using 3D deformable model and camera self-calibration," *arXiv preprint arXiv:1708.06831*, 2017.

[22] S. R. Ke, L. Zhu, J.-N. Hwang, H.-I. Pai, K.-M., Lan and C. P. Liao, "Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming," in *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, pp. 489-496), 2010.

[23] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, 1964.

[24] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, 2014.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP J. Image and Video Processing*, *2008*.

[26] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Information Processing Sys.*, pp. 91-99, 2015.

[27] Z. Tang, G. Wang, H. Xiao, A. Zheng and J.-N. Hwang, "Single-camera and inter-camera vehicle tracking and 3D speed estimation based on fusion of visual and semantic features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. Workshops*, 2018.