Cross-View Tracking for Multi-Human 3D Pose Estimation at over 100 FPS Supplementary material

Long Chen[1] Haizhou Ai[1] Rui Chen[1] Zijie Zhuang[1] Shuang Liu[2]
[1]Department of Computer Science and Technology, Tsinghua University [2]AiFi Inc.

1. Detail of Target Initialization

Here, we present details of our target initialization algorithm, including the epipolar constraint, cycle-consistency, and the formulation we utilized for graph partitioning.

When two cameras observing a 3D point from two distinct views, the epipolar constraint [2] provides relations between the two projected 2D points in camera coordinates, as illustrated in Figure 1. Supposing $\mathbf{x}_L$ is the projected 2D point in the left view, the another projected point $\mathbf{x}_R$ of the right view should be contained in the epipolar line:

$$l_R = \mathrm{F}\mathbf{x}_L, \tag{1}$$

where F is the fundamental matrix that determined by the internal parameters and relative poses of the two cameras. Therefore, given two points from two views, we can measure the correspondence between them based on the point-to-line distance in the camera coordinates:

$$A_e(\mathbf{x}_L, \mathbf{x}_R) = 1 - \frac{d_l(\mathbf{x}_L, l_L) + d_l(\mathbf{x}_R, l_R)}{2 \cdot \alpha_{2D}}. \tag{2}$$

Given a set of unmatched detections $\{D_i\}$ from different cameras, we compute the affinity matrix using Equation 2. Then the problem is turned to associate these detections across camera views. Note that there are multiple cameras, the association problem can not be formulated as simple bipartite graph partitioning. And the matching result should satisfy the cycle-consistent constraint, i.e. $\langle D_i, D_k \rangle$ must be matched if $\langle D_i, D_j \rangle$ and $\langle D_j, D_k \rangle$ are matched. To this end, we formulate the problem as general graph partitioning and solve it via binary integer programming [1, 3]:

$$\mathbf{y}^* = \underset{Y}{\operatorname{argmax}} \sum_{\langle D_i, D_j \rangle} a_{ij} y_{ij}, \tag{3}$$

subject to

$$y_{ij} \in \{0, 1\}, \tag{4}$$

$$y_{ij} + y_{jk} \leq 1 + y_{ik}, \tag{5}$$

where a_{ij} is the affinity between $\langle D_i, D_j \rangle$ and Y is the set of all possible assignments to the binary variables y_{ij}. The cycle-consistency constraint is ensured by Equation 5.

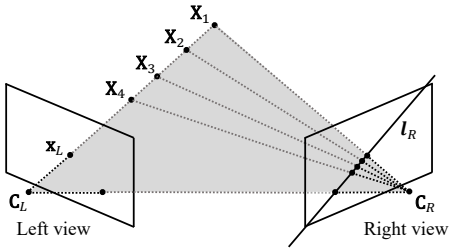

Figure 1: Epipolar constraint: given $\mathbf{x}_L$, the projection on the right camera plane $\mathbf{x}_R$ must be on the epipolar line l_R.

2. Baseline Method in the Ablation Study

To verify the effectiveness of our solution, we construct a method that matches joints in pairs of views using epipolar constraint as the baseline in ablation study. The procedure of the baseline method is detailed in Algorithm 1. Basically, for each frame, it takes 2D poses from all cameras as inputs, and associate them across views using epipolar constraint and graph partitioning. Afterwards, 3D poses are estimated from the matching results via triangulation.

3. Parameter Selection

In this work, we have six parameters: w_{2D}, w_{3D} are the weights of the affinity measurements, α_{2D} and α_{3D} are the corresponding thresholds, and λ_a, λ_t are the time penalty rates for the affinity calculation and incremental triangulation, respectively. Here in Table 1, we first show the experimental results with different affinity weights on the Campus dataset. As seen in the table, 3D correspondence is critical in our framework but the performance is robust to the combination of weights. Therefore, we fix $w_{2D} = 0.4$, $w_{3D} = 0.6$ for all datasets, and select other parameters for each dataset empirically, as shown in Table 2. The basic intuition behind it is to adjust α_{2D} according to the image resolution and change λ_a, λ_t based on the input frame rate. Since different datasets are captured at different frame rates, e.g. the first three public datasets are captured at 25 FPS

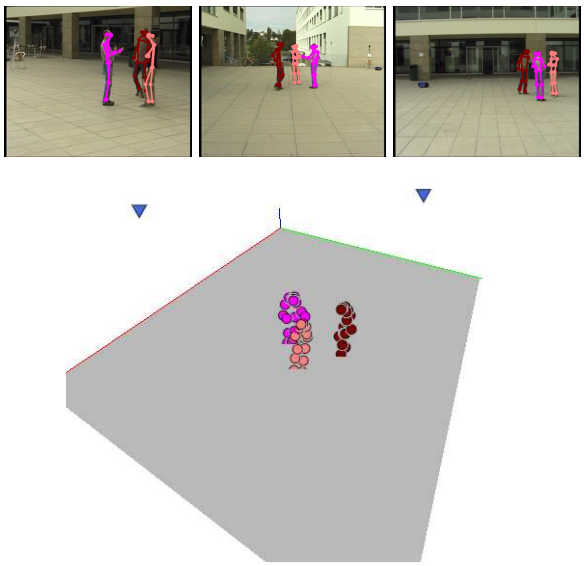

Figure 2: Qualitative result on the Campus dataset. There are three people with three cameras in an outdoor square. Different people are represented in different colors based on the tracking result. The camera locations are illustrated in the 3D view as triangles in blue.

while the Store dataset is captured at 10 FPS.

4. Qualitative Results

Here, we present more qualitative results of our solution on public datasets in Figure 2, Figure 3, and Figure 4. A recorded video is also provided at https://youtu.be/-4wTcGjHZq8, to demonstrate the quality of our method in multi-view 3D pose estimation and multi-human tracking.

Algorithm 1: Baseline for 3D pose estimation.

Input: 2D human poses $\mathbb{D} = \{D_{i,c_i} | i = 1, ..., M\}$
Output: 3D poses of all people $\mathbb{T} = \{T_i\}$
Initialization: $\mathbb{T} \leftarrow \emptyset$; $\mathbf{A} \leftarrow \mathbf{A}_{M \times M} \in \mathbb{R}^{M \times M}$
foreach $D_{i,c_i} \in \mathbb{D}$ **do**
  foreach $D_{j,c_j} \in \mathbb{D}$ **do**
    if $c_i \neq c_j$ **then**
      $\mathbf{A}(i,j) \leftarrow A_e(D_{i,c_i}, D_{j,c_j})$
    else
      $\mathbf{A}(i,j) \leftarrow -\inf$
    end
  end
end
foreach $\mathbb{D}_{cluster} \in \text{GraphPartitioning}(\mathbf{A})$ **do**
  if $\text{Length}(\mathbb{D}_{cluster}) \geq 2$ **then**
    $\mathbb{T} \leftarrow \mathbb{T} \cup \text{Triangulation}(\mathbb{D}_{cluster})$
  end
end

w_{2D}	w_{3D}	Association Accuracy (%)	PCP (%)
1.0	0.0	45.69	62.29
0.8	0.2	96.22	96.58
0.6	0.4	96.30	96.61
0.4	**0.6**	**96.38**	**96.63**
0.2	**0.8**	**96.38**	**96.63**
0.0	1.0	96.38	96.49

Table 1: Association accuracy and PCP score with different weight combinations on the Campus dataset.

Dataset	α_{2D} *(pixel / second)*	α_{3D} *(m)*	λ_a	λ_t
Campus	25	0.10	5	10
Shelf	60	0.15	5	10
CMU Panoptic	60	0.15	5	10
Store (layout 1)	70	0.25	3	5
Store (layout 2)	70	0.25	3	5

Table 2: Parameter selection for each dataset.

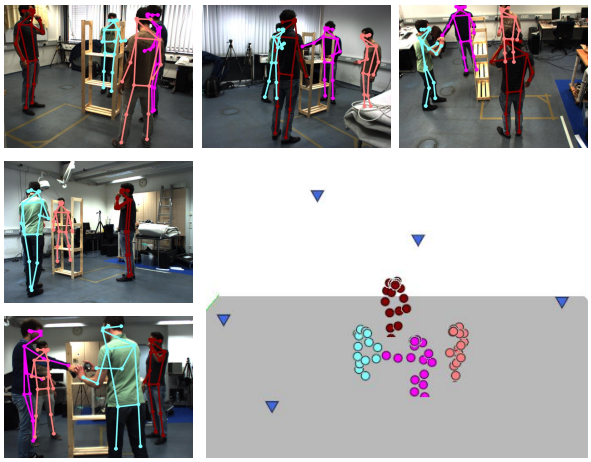

Figure 3: Qualitative result on the Shelf dataset. It consists of four people disassembling a shelf under five cameras. The camera locations are illustrated in the 3D view as triangles in blue. The actions of people can be seen clearly from the estimated 3D poses.

References

[1] Long Chen, Haizhou Ai, Rui Chen, and Zijie Zhuang. Aggregate tracklet appearance features for multi-object tracking. *IEEE Signal Processing Letters*, 26(11):1613–1617, 2019. 1

[2] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[3] Ergys Ristani and Carlo Tomasi. Tracking multiple people online and in real time. In *Asian conference on computer vision*, pages 444–459. Springer, 2014. 1

Figure 4: Qualitative result on the CMU Panoptic dataset. There are 31 cameras and 7 people in the scene. The cameras are distributed over the surface of a geodesic sphere. As we detailed in ablation study, with the proposed iterative processing all the 31 cameras can be updated in around 0.058 seconds.