Efficient Spatio-Temporal Data Association Using Multidimensional Assignment for Multi-Camera Multi-Target Tracking

Moonsub Byeon
msbyeon@snu.ac.kr
Songhwai Oh
songhwai@snu.ac.kr
Kikyung Kim
koreaton@snu.ac.kr
Haan-Ju Yoo
neohanju@snu.ac.kr
Jin Young Choi
jychoi@snu.ac.kr

Perception and Intelligence Laboratory,
Department of
Electrical and Computer Engineering,
ASRI, Seoul National University,
Seoul, Korea

Abstract

This paper proposes a novel multi-target tracking method which jointly solves a data association problem using images from multiple cameras. In this work, the spatio-temporal data association problem is formulated as a multidimensional assignment problem (MDA). To achieve a fast, efficient, and easily implementable approximation algorithm, we solve the MDA problem approximately by solving a sequence of bipartite matching problems using random splitting and merging operations. In this formulation, we design a new cost function, considering the accuracy in 3D reconstruction, motion smoothness, visibility from cameras, starting/ending at entrance and exit zone, and false positive. Our approach reconstructs 3D trajectories that represent people's movement as 3D cylinders whose locations are estimated considering all adjacent frames. The experiments illustrate the proposed method shows the state-of-the-art performance in challenging multi-camera datasets and the computational efficiency with 8 times faster computation than the existing BIP approach.

1 Introduction

3D localization and tracking of multiple targets are important tasks in computer vision for applications such as surveillance and sports player analysis. A number of multi-target tracking algorithms have been proposed for many years giving successful results [3, 10, 11, 12, 13, 16, 19, 20]. In single camera approaches [3, 20], 3D locations of a person is estimated with a ground plane assumption that the person stands on a 3D virtual plane. Most of the recent single camera approaches adopt the tracking-by-detection framework which utilizes time-independent observations obtained by the classification or background subtraction methods. In the tracking-by-detection framework, tracking means linking the observations through

frames for the same person, so the problem is referred to as *data association*. The benefit of tracking-by-detection is that it is robust to drifting and easy to recover from tracking failure. However, occlusion makes the problem still challenging because the detection performance significantly decreases when people are occluded by static obstacles or other people. To overcome the occlusions, the multi-camera based approaches have attempt to integrate the observations from multi-cameras. Several multi-camera based approaches have generated a probability map [10, 12, 19]. After reconstructing these probability maps, they apply a single camera based tracking approach, such as linear programming [1], graph cut [12], and mean-shift single target tracker [19]. Some other approaches have made 3D hypotheses by fusing object detections from multi-cameras and solved (temporal) data association problem of the 3D hypotheses [4, 21]. The main disadvantage of these multi-camera based approaches is that they separate the problem into two sub-problems: reconstruction and tracking.

In recent years, there has been an increasing interest in a unified framework considering reconstruction and tracking simultaneously [11, 13]. In the unified framework, two combinatorial problems should be solved at the same time: spatial data association between cameras and temporal data association between frames. Since the *spatio-temporal data association* problem is a well known NP-hard problem even in a small number of cameras or frames (more than 3) [18], it is difficult to make the problem tractable. Recently, several min-cost flow based methods [11, 13] formulated the spatio-temporal data association problem as a binary integer programming (BIP) problem generating a graph among detections and solved it by a BIP solver. To make the BIP problem tractable, they simplify both optimization variables and the cost function by assuming that the optimization variables and the cost function are decomposable with respect to edges of the graph. Although several min-cost flow based methods in a single camera can get a global optimum in a polynomial time [1, 22], the min-cost flow based methods in multi-camera are still NP-hard. The complexity grows exponentially with the number of cameras since combinations of observations from different cameras exponentially increase and their costs need to be predefined for a BIP solver. In addition, the min-cost flow based approaches can not deal with a higher-order motion model which is a function of three or more nodes because the cost function of the min-cost flow approaches can not be factored into the product/sum of edges of multiple adjacent nodes.

In this paper, we propose an approximation algorithm for multidimensional assignment problem to solve the spatio-temporal data association problem with a reasonable computational load. The approximation algorithm iteratively improves a feasible solution by two operations: random splitting and merging. The solution in each iteration is re-constructed by random splitting and optimal merging of the trajectories in the previous solution. The new solution is evaluated by the proposed cost function and obtained so as to have lower cost than the previous one and eventually the solution converges to the local minimum. Given a feasible solution, the approximation problem corresponds to a bipartite matching problem of random splits of the previous set of trajectories. Hence the proposed formulation can be considered as a guided random search to find the global optimum through repeated random local searches (bipartite matchings). In addition, we design a new cost function considering 3D reconstruction accuracy, motion smoothness, visibility from cameras, starting/ending at entrance and exit zone, and false positive. In particular, we use a spline-based probability model for higher-order motion model to improve the tracking performance by considering the dynamic patterns of motions. We also pursue a high accuracy estimation of 3D trajectories by smoothing the adjacent 3D positions. To evaluate the performance of 3D trajectory estimation, we present a new dataset containing the ground truth of 3D head trajectories of each person. The experiments using the dataset illustrate our 3D trajectory reconstruction

and higher-order motion model significantly improves the 3D tracking performance. Additional experiments are conducted on public multi-camera datasets to compare our method with the state-of-the-arts in view of performance and computation.

2 Proposed Method

We adopt the tracking-by-detection framework which considers every object detection as an observation. A set of detections from all cameras is denoted by $\mathbf{D}$ whose element $\mathbf{d}_i^{k,t} \in \mathbf{D}$ represents 2D bounding box for the i-th detection at frame index $t \in \{1,...,T\}$ and camera index $k \in \{1,...K\}$. We have m_{kt} observations at the t-th frame of the k-th camera and each observation is labeled by an augmented index set

$$\mathbf{I}_{kt} = \{0,1,2,...,m_{kt}\}; \tag{1}$$

where dummy index 0 represents a missing or invisible detection. The observations during $1,...,T$ frames and at cameras $1,...K$ form a KT-partite (hyper) graph,

$$G = (V,E) = (\mathbf{I}_{11} \cup ... \cup \mathbf{I}_{KT}, E), \tag{2}$$

where vertices V are partitioned into $K \times T$ different independent sets $\mathbf{I}_{11},...,\mathbf{I}_{KT}$ and each hyperedge in E contains at least one vertex in each partite set.

Trajectory hypotheses set $\mathbf{T}$ can be defined as a set of all hyperedges E. We represent each trajectory hypothesis $\mathcal{T}_n \in \mathbf{T}$ as a matrix whose entry in the k-th row and t-th column corresponds to an observation index at the t-th frame of the k-the camera. In $K = 3$, $T = 5$ case, for example, a trajectory hypothesis can be expressed by

$$\mathcal{T}_n = \begin{pmatrix} i_{11}^n & i_{12}^n & i_{13}^n & i_{14}^n & i_{15}^n \\ i_{21}^n & i_{22}^n & i_{23}^n & i_{24}^n & i_{25}^n \\ i_{31}^n & i_{32}^n & i_{33}^n & i_{34}^n & i_{35}^n \end{pmatrix}, i_{kt}^n \in \mathbf{I}_{kt}. \tag{3}$$

2.1 Problem Formulation

Our goal is to find the optimal association hypothesis representing the true trajectories of all persons captured by all cameras, where the association hypothesis $\mathbf{H}$ is denoted by a set of disjoint trajectory hypotheses $\mathcal{T}_n$, $n = 1,...,p$ with unknown number p of persons, that is, $\mathbf{H} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_p\}$. Each trajectory hypothesis $\mathcal{T}_n$ corresponds to a single person and it is assumed that each trajectory hypothesis does not share an observation with other trajectory hypotheses, i.e.,$\forall i_{kt}^n, i_{kt}^m \neq 0$, $i_{kt}^n \neq i_{kt}^m$ if $n \neq m$. Our data association problem is achieved by minimizing the sum of costs of trajectories over the association hypothesis $\mathbf{H}$.

The problem of finding a set of disjoint trajectory hypotheses with a minimum sum of costs can be formulated as a multidimensional assignment problem which is equivalent to the problem of minimizing the sum of costs of hyperedges containing one element per partite set in the hypergraph G. With binary decision variables $x_{\mathcal{T}_n} \in \{0,1\}$ deciding whether the trajectory $\mathcal{T}_n$ is in the association hypothesis $\mathbf{H}$, and cost function $c : \mathbf{T} \to \mathbb{R}$, the objective function and disjointness constraints are given by

$$\min \sum_{\mathcal{T}_n \in \mathbf{T}} c(\mathcal{T}_n) x_{\mathcal{T}_n} \quad s.t. \sum_{\mathcal{T}_n \in \mathbf{T}_{[kt],i}} x_{\mathcal{T}_n} = 1 \quad \begin{cases} \mathbf{T}_{[kt],i} = \{\mathcal{T}_n \in \mathbf{T} | [\mathcal{T}_n]_{k,t} = i\} \\ k = 1,2,...,K \\ t = 1,2,...T \\ i = 1,2,...,m_{kt} \end{cases} \tag{4}$$

where $\mathbf{T}_{[kt],i}$ is a subset of all trajectory hypothesis set $\mathbf{T}$ whose value in the k-th row and t-th column has the detection index i and cost $c(\mathcal{T}_n)$ denotes the cost of the n-th trajectory $\mathcal{T}_n$. Because of combinatorial space of $\mathbf{T}$ and disjoint constraints, the problem becomes NP-hard even in small T or K (more than 3 cameras or 3 frames) [18]. It can be solved by using an approximate method such as greedy, branch and bound techniques, and the Lagrangian relaxation methods [7, 18]. In a single camera case, Collins [5] proposed an approximate algorithm that iteratively improve a feasible solution to solve multi-target tracking problem. Until now, there is no research on the MDA formulation to solve both the multi-target tracking problem and the multi-camera fusion problem simultaneously. The following section presents an algorithm for finding an approximate solution to the MDA problem by iteratively improving the given initial solution.

2.2 Approximation of MDA

If MDA problem can be approximated to a weighted bipartite matching problem, it can be solved in a polynomial time by a well known algorithm such as Hungarian algorithm [17]. In this section, we formulate a weighted bipartite matching problem approximating the original MDA formulation eq. (4). Denoting an association hypothesis at $iter$-th iteration as $\mathbf{H}^{iter}$, we can get a new association hypothesis $\mathbf{H}^{iter}$ by improving the previous association hypothesis $\mathbf{H}^{iter-1}$. The key idea is that we randomly split the previous association hypothesis $\mathbf{H}^{iter-1}$ and optimally re-merging by solving the approximated matching problem. Our split/merge strategy is expected to find the solution of which quality is better than or equal to the previous one, because it splits the previous association hypothesis $\mathbf{H}^{iter-1}$ while maintaining the disjointness of the trajectory and re-merges the split trajectories to form a new association hypothesis $\mathbf{H}^{iter}$ with a less than or equal to the previous cost.

First, we randomly split the previous association hypothesis $\mathbf{H}^{iter-1}$ into two groups $\widetilde{\mathbf{H}}^I$ and $\widetilde{\mathbf{H}}^J$. With the matrix representation of trajectory hypothesis, a single trajectory hypothesis can be divided into two splits by using two binary matrices called Random Split Mask (RSM) satisfying

$$\mathcal{M}^I + \mathcal{M}^J = \mathbb{1}^{T\times K}, \tag{5}$$

where $\mathbb{1}^{T\times K}$ is all-ones matrix and all entries of $\mathcal{M}^I, \mathcal{M}^J$ have 0 or 1. With this Random Split Mask $\mathcal{M}^I, \mathcal{M}^J$, one trajectory hypothesis $\mathcal{T}_n$ can be divided into

$$\mathcal{T}_n = \mathcal{M}^I \otimes \mathcal{T}_n + \mathcal{M}^J \otimes \mathcal{T}_n, \tag{6}$$

where $\otimes$ is an entry-wise product operator. $\mathcal{M}^I \otimes \mathcal{T}_n$ and $\mathcal{M}^J \otimes \mathcal{T}_n$ are trajectory hypothesis splits. Finally, an association hypothesis $\mathbf{H}^{iter-1}$ is split into $\widetilde{\mathbf{H}}^I$ and $\widetilde{\mathbf{H}}^J$ as follows:

$$\widetilde{\mathbf{H}}^I = \{\mathcal{M}^I \otimes \mathcal{T}_n | \quad \forall \mathcal{T}_n \in \mathbf{H}^{iter-1}, \mathcal{M}^I \otimes \mathcal{T}_n \neq 0\} \tag{7}$$

$$\widetilde{\mathbf{H}}^J = \{\mathcal{M}^J \otimes \mathcal{T}_n | \quad \forall \mathcal{T}_n \in \mathbf{H}^{iter-1}, \mathcal{M}^J \otimes \mathcal{T}_n \neq 0\}. \tag{8}$$

Merging two trajectory hypothesis can be interpreted as a summation of the re-selected two trajectory hypothesis splits from different trajectory hypotheses. For each $\mathcal{T}_i^I \in \widetilde{\mathbf{H}}^I, \mathcal{T}_j^J \in \widetilde{\mathbf{H}}^J$, the merged trajectory hypothesis is the summation of two matrices $\mathcal{T}_i^I + \mathcal{T}_j^J$. Next, we are going to find an optimal merging pair $\mathcal{T}_i^I, \mathcal{T}_j^J$, which is equivalent to the problem of finding a weighted maximum matching in a bipartite graph. The objective function of merging process

is formulated as

$$\min \sum_{i=0}^{|\widetilde{\mathbf{H}}^I|} \sum_{j=0}^{|\widetilde{\mathbf{H}}^J|} \tilde{c}_{ij} \psi_{ij} \quad s.t. \quad \begin{array}{l} \sum_{j=0}^{|\widetilde{\mathbf{H}}^J|} \psi_{ij} = 1; \quad i = 1,...,|\widetilde{\mathbf{H}}^I| \\ \sum_{i=0}^{|\widetilde{\mathbf{H}}^I|} \psi_{ij} = 1; \quad j = 1,...,|\widetilde{\mathbf{H}}^J|, \end{array} \tag{9}$$

where $\tilde{c}_{ij} = c(\mathcal{T}_i^I + \mathcal{T}_j^J)$ and $\psi_{ij} = 1$ means a merging of two trajectory hypotheses $\mathcal{T}_i^I$ and $\mathcal{T}_j^J$. The problem can be solved exactly in polynomial time by the Kuhn-Munkres Hungrarian algorithm [17]. From the optimal solution ψ_{ij}^* of the weighted bipartite matching problem, the re-merged trajectory hypothesis $\mathcal{T}_n \in \mathbf{H}^{iter}$ is obtained by $\mathcal{T}_n = \mathcal{T}_i^I + \mathcal{T}_j^J, \forall i, j, \forall \psi_{ij}^* = 1$. Finally, the new association hypothesis $\mathbf{H}^{iter}$ is a set of re-merged trajectory between $\widetilde{\mathbf{H}}^I$ and $\widetilde{\mathbf{H}}^J$.

In our data association, the optimal value C^{iter} of the new association hypothesis $\mathbf{H}^{iter}$ at the *iter*-th iteration can not be worse than the previous optimal value C^{iter-1}, that is $C^{iter} \leq C^{iter-1}$. This is because we adopt a descent strategy, in other words, the association hypothesis would not change when the new association has larger optimal value than the previous association. Finally, it is guaranteed to eventually converged to a local optimum by randomly changing the RSM. We have empirically found that the proposed algorithm converges rapidly. Our optimization strategy is summarized in Algorithm 1.

Our method finds the optimum by iteratively improving a feasible solution, so the initial feasible solution is required. We first find a possible association in spatial domain, considering the distance between detections in world coordinate. In our experiment, the maximum of allowed distance is set to 1.5 meter. Next, we find the detections with the minimum reconstruction error at each frame, then we regard the detections as a person candidate for the next temporal data association. We iteratively find the other person candidates with minimum reconstruction error until all detections are selected. Next, we link the candidates in temporal domain. We sequentially solve a bipartite matching problem between two frames by the Hungarian algorithm and define weights of each edge as distances between the candidates. For the initial feasible solution, we note that only reliable candidates are linked. We ignore all edges from a candidate if the ratio of the the second smallest weight of edge from the candidate to the smallest one is smaller than a threshold $\tau = 1.5$. We use this greedy method for baselines of our experiment, denoted as "Initial" and "Greedy" method, but in "Greedy" method consider all edges between candidates without ignoring them for long trajectories.

3 Cost Design

In this section, we present a cost design for our formulation, considering 3D reconstruction accuracy, motion smoothness, starting/ending at entrance and exit zone, and false positive. We model each person as a 3D cylinder centered at (x,y,z) in a 3D space with radius r and height h. We assume that there exists a deterministic function mapping a trajectory hypothesis $\mathcal{T}_n$ to a set of 3D cylinder $\mathbf{X}_n = \{\mathbf{x}_n^{s_n}, ..., \mathbf{x}_n^{e_n}\}$ where each s_n and e_n, respectively, are index of the first and last frames. $\mathbf{x}_n^t$ is a 3D cylinder at the t-th frame. The set of detections of the n-th person at the t-th frame is denoted by $\mathbf{D}_n^t$ and the index set of visible cameras of $\mathbf{X}_n^t$ is denoted by $\mathbf{N}_n^t$.

Algorithm 1 An Iterative Approximation of MDA

Input: $\mathbf{H}^0$, *max_iteration*
Output: $\mathbf{H}^*$

```
for iter ← 1,...,max_iteration do
    Select the Random Split Mask M^I, M^J
    Split H^{iter-1} into H̃^I, H̃^J by RSM M^I, M^J
    H̃^I ← {M^I ⊗ T_i}, ∀T_i ∈ H^{iter-1}
    H̃^J ← {M^J ⊗ T_i}, ∀T_i ∈ H^{iter-1}
    Find optimal weighted bipartite matching ψ*_ij
    for all i, j satifying ψ*_ij = 1 do
        H^{iter} ← T^I_i + T^I_j
    end for
end for
H* = H^{max_iteration}
```

Our cost function is a summation of five individual terms: cost for 3D reconstruction accuracy (c_{rec}), cost for motion smoothness (c_{mot}), cost for visibility from camera (c_{vis}), cost for trajectory start/end (c_{tse}), and cost for false positive trajectory (c_{fpt}),

$$c(\mathcal{T}_n) = c_{\mathcal{T}_n} = c_{rec} + c_{mot} + c_{vis} + c_{tse} + c_{fpt}. \tag{10}$$

Cost for 3D Reconstruction Accuracy. We design the cost c_{rec} measuring 3D reconstruction error between a 3D cylinder model and its observations from multi-cameras. The cost c_{rec} increases proportional to the Euclidean distances between the 3D cylinder and projections from 2D camera observations. We define the c_{rec} as a summation of 3D reconstruction error ε_{rec} over the entire trajectory. At each frame, 3D reconstruction error ε_{rec} is the root mean squared error (RMSE) between the center of 3D cylinder $\mathbf{c}_n^t = (x, y, z)$ and projections from the detection set $\mathbf{D}_n^t$. We also regularize the 3D reconstruction error by setting a default error term r when a person is visible but not detected. The cost for the 3D reconstruction accuracy is given by

$$c_{rec}(\mathbf{X}_n) = \sum_{t=s_n}^{e_n} \lambda_{rec} \cdot \varepsilon_{rec}(\mathbf{x}_n^t) = \sum_{t=s_n}^{e_n} \lambda_{rec} \cdot \sqrt{\frac{1}{|\mathbf{N}_n^t|} \sum_{k \in \mathbf{N}_n^t} \varepsilon_{rec}^k(\mathbf{X}_n^t)^2} \tag{11}$$

$$\varepsilon_{rec}^k(\mathbf{x}_n^t) = \begin{cases} \left\| \Phi^k(\mathbf{d}_n^{k,t}, z) - \mathbf{c}_n^t \right\|, & \exists \mathbf{d}_n^{k,t} \in \mathbf{D}_n^t \\ r, & otherwise \end{cases} \tag{12}$$

where $\Phi^k(\mathbf{d}, z)$ is the projected point of detection $\mathbf{d}$ from the image plane of k-th camera to the world coordinate where z-coordinate is fixed to z.

Cost for Motion Smoothness. The cost for motion smoothness evaluate how well trajectory describes real motion of a person. We assume that motion of people is closer to a higher-order curve rather than a first-order line. For this purpose, we adopt a spline-based cost function for higher-order motion model considering curvature term ε_c of a curve as well as average distance ε_d similar to Collins [5]. The cost is defined by a weighted sum of curvature term ε_c and average distance term ε_d and the weights are given by the number of average detections over the consecutive frames. The cost for motion smoothness is given by

$$c_{mot}(\mathbf{X}_n) = \lambda_{mot} \cdot (\alpha_m \cdot \varepsilon_d + (1 - \alpha_m) \cdot \varepsilon_c). \tag{13}$$

where each term is given by $\varepsilon_d = \sum_{t=2}^{e_n} w_d^t \cdot \left\| \mathbf{x}_n^t - \mathbf{x}_n^{t-1} \right\|, \varepsilon_c = \sum_{t=2}^{e_n-1} w_c^t \cdot \left\| \mathbf{x}_n^{t+1} - 2 \cdot \mathbf{x}_n^t + \mathbf{x}_n^{t-1} \right\|$, $w_d^t = \frac{|\mathbf{D}_n^t| + |\mathbf{D}_n^{t-1}|}{2}$, and $w_c^t = \frac{|\mathbf{D}_n^{t+1}| + |\mathbf{D}_n^t| + |\mathbf{D}_n^{t-1}|}{3}$.

Cost for Visibility from Cameras The cost for visibility of cameras is designed for modeling whether a person is visible at each camera. We assume that a person will be detected at each camera with probability of recall α_{rec}. The cost for visibility from camera increases proportional to the number of cameras from which a person is visible but not detected. We give a penalty to the trajectory which has missing detections, satisfying

$$c_{vis}(\mathbf{X}_n) = \sum_{t=s_n}^{e_n} (|\mathbf{N}_n^t| - |\mathbf{D}_n^t|) \cdot \log(1 - \alpha_{rec}), \tag{14}$$

Cost on Trajectory Start/End This cost is to prevent a trivial solution which includes too many track fragments. The long trajectory hypothesis is more reasonable than the short one. Similar to single camera approach [22], we give a penalty to the trajectory hypothesis whenever it starts or ends, which means that a trajectory with frequent starting or ending increases the cost. The cost is proportional to the number of detection at start/end time $c_{tse} = \lambda_{tse} \cdot (|\mathbf{D}_n^{s_n}| + |\mathbf{D}_n^{e_n}|)$. In addition, a trajectory hypothesis is enforced to start and end at entrance and exit zone. If a trajectory hypothesis starts or ends at a region out of the entry/exit zone, we give an additional penalty proportional to the number of detections, that is, $\lambda_{tee} \cdot |\mathbf{D}_n^t|$ where λ_{tee} is a design parameter.

Cost for False Positive Trajectory The cost for false positive trajectories prevents a trivial association hypothesis where a trajectory hypothesis is considered as a false positive. We penalize a false positive trajectory which consists of detections at the frame where a trajectory starts and ends at the same time. The cost can be defined as $c_{fpt}(\mathcal{T}_n) = \lambda_{fpt} * |\mathbf{D}_n^{s_n}|$, if $s_n = e_n$.

4 Experimental Results

We evaluated the performance of our 3D localization and tracking method on several challenging datasets. Our goal is to provide the 3D locations of heads. However, there is no standard dataset for evaluating 3D locations of heads, so we have constructed a new dataset which provides the ground truth of 3D trajectories. For quantitative evaluation, we used the widely accepted CLEAR MOT metrics [2]. The Multi-Object Tracking Accuracy (MOTA) measures the overall performance of multi-target tracking considering missed targets, false alarms, and identity switches. The Multi-Object Tracking Precision (MOTP) averages the localization error between estimated and ground truth trajectory. We applied CLEAR MOT metrics [2] in 3D world coordinate. The criterion on matching to the ground truth is defined as the distance in 3D world coordinate, where the target is determined to be matched to the ground truth when the distance is less than a threshold (set to 1 meter in our experiment). Moreover, we also used the metrics proposed in [15], that is, the number of trajectories mostly tracked (MT), mostly lost (ML), and partially tracked (PT) as well as the number of fragments (FM) and identity switches (IDS). We reported the numbers for recall and precision. Figure 2 illustrates our tracking results. In our all experiments, we set $r = 30$ cm, $\alpha_{rec} = 0.9$ and $\alpha_m = 0.5$ which indicates the diameter of a person, the expected recall of the detector, and a weight of motion smoothness. People are expected to move under the

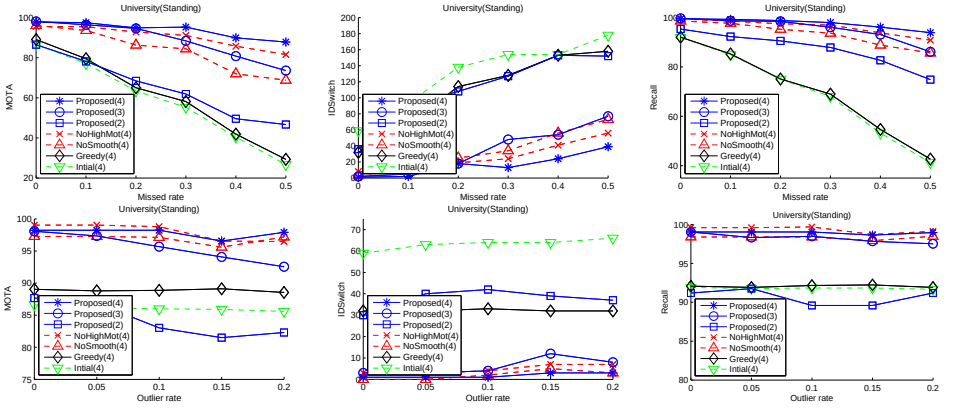

Figure 1: Results for MOTA, IDS and Recall in PSN-University *standing* sequence. Top: for increasing missing data. Bottom: for increasing false positives. Also shown in parentheses is the number of cameras.

maximum allowed distance $\alpha_{max} = 840/f$ cm at a video with average frame-rate f fps. We set $(\lambda_{rec}, \lambda_{mot}, \lambda_{tse}, \lambda_{tee}, \lambda_{fpt}) = (0.5, 1/\alpha_{max}, 5, 3.34, 12.5)$ through trial and error method.

PSN-University. The PSN-University dataset consists of three sequences of *standing*, *sitting*, and *standing & sitting*. The video was captured by four 10 Megapixel cameras with 3648×2752 resolution at 3 fps. In our experiment, we did not use any detectors to decouple the tracking performance from the detector's performance, that is, we used a video with hand-labeled head 2D bounding boxes. Instead, to evaluate the tracking performance depending on the detector's performance, we added missing data and outliers to the annotated ground truth detections. To simulate missing data, we removed true detections randomly, where we changed the missing rate from 0% to 50%. To create the outliers, we made false detections at random locations chosen uniformly, where we changed the outlier rate from 0 to 20% of true detections

Figure 1 shows tendency of MOTA, IDS, and Recall depending on the missing rate and outliers rate. The result shows that both "Initial" and "Greedy" method (details in Section 2.2) in 0% missing rate and 0% outlier rate achieved over 85% MOTA, but tracking performance is significantly degraded for increasing detector errors. In Figure 1, we report the results for different number of cameras as denoted blue line. We observed that adding cameras results in significant improvement in performance for both missing data and outliers case. MOTA of our method with four cameras drops only 10% for the missing rate of 50%, while MOTA with two cameras drops 40%. Similarly, for the 20% outliers, MOTA with two cameras drops 20% more than that with four cameras. Therefore, we can conclude that using more cameras yields more robust performance against detector's errors. To validate the effectiveness of our key cost functions, we also report the results of the variants of our method denoted red dash lines: without higher-order motion model (NoHighMot), without smoothing adjacent 3D positions (NoSmooth). The result implies that the proposed higher-order motion model and smoothing adjacent 3D positions improve overall tracking performance in terms of MOTA, IDS, and Recall.

We further report the results when both missing data and outliers are contained simultaneously. Table 1 shows the results of all sequences of the PSN-University dataset with 15% outliers and 30% missing data which is the usual case in most detectors. *sitting* and *stand-*

Dataset	Method	MOTA	MOTP	IDS	FM	MT	PT
PSN-University *standing* seq.	**Our method***	89.0	77.1	19	18	9	1
	Our method	83.3	75.5	26	20	10	0
	NoHighMot	82.5	73.0	32	23	10	0
	NoSmooth	73.0	72.7	43	40	9	1
	Greedy	53.1	74.3	141	133	0	10
PSN-University *sitting* seq.	**Our method***	85.9	86.9	15	11	10	0
	Our method	77.6	88.3	36	36	9	1
	NoHighMot	75.4	84.7	57	45	10	0
	NoSmooth	61.3	80.6	86	74	8	2
	Greedy	52.7	83.4	203	200	2	8
PSN-University *sit.&stand.* seq.	**Our method***	78.3	84.7	28	20	9	1
	Our method	85.6	89.4	32	19	10	0
	NoHighMot	87.6	87.0	29	31	10	0
	NoSmooth	67.3	81.6	111	94	8	2
	Greedy	58.9	85.4	268	263	5	5

Table 1: Quantitative result on all sequences of the PSN-University dataset. * mark denotes the results when used with current existing classifier [8].

Method	MOTA	MOTP	FM	IDS	Rcll	Prcn
Ours (3)	98.5	77.5	2	2	99.2	99.3
Ours (5)	**99.4**	77.8	**0**	**0**	**99.6**	**99.7**
[1] (5)	82	50	-	-	-	-
[13] (2)	76.0	60	-	-	-	-
[13] (3)	71.4	53.4	-	-	-	-
[11] (3)	**99.4**	**83.0**	1	2	-	-

Table 2: Quantitative result on PETS2009. The number of cameras is shown in parentheses.

ing & sitting sequence have a large variation of heights of people because there are people sitting and standing. In *standing & sitting* sequence, MOTA rises by about 18% when using reconstruction of trajectories considering all adjacent frames. This is because considering adjacent frames can estimate more reasonable 3D trajectories when a person is detected by a single camera. We also report the results when used with current existing detector, denoted as a star mark. We trained the state-of-the-arts pedestrian classifier [8] for a head detector using positive samples from [14] and negative samples from INRIA dataset [6].

PETS2009. We evaluated our method on PETS2009 dataset and compared our result to the state-of-the-arts Berclaz et al.[1], Leal-Taixe *et al.*[13] and Hoffmann *et al.* [11] as shown in Table 2. Similar to Hoffmann *et al.* [11], we used the deformable part model (DPM) [9] for the detection of people and adopted the same 3D ground truth trajectories provided by Milan *et al.* [16]. Note that our results are much better than the proposed ones in [1, 13]. Compared with the current best-performing method [11], our method also reported the lowest ID switches (IDS), fragments (FM), and competitive performance on MOTA measure by using five cameras. We achieved high recall and precision, which means most of people are localized and tracked with consistent identities. For this dataset, we also evaluated the influence of the number of cameras. As in the case of the PSN-University dataset, the increase of cameras improves the performance of our method, while the performance in [13] is degraded as the number of cameras increases.

To show the efficiency of our algorithm, we compared computational time of our method with that of the best-performing method [11] on PETS2009 dataset. For a fair comparison, computational time includes both time for calculating predefined costs and time for a BIP solver in [11]. When we used five cameras and performed on an i7 CPU, 3.4 GHz, and

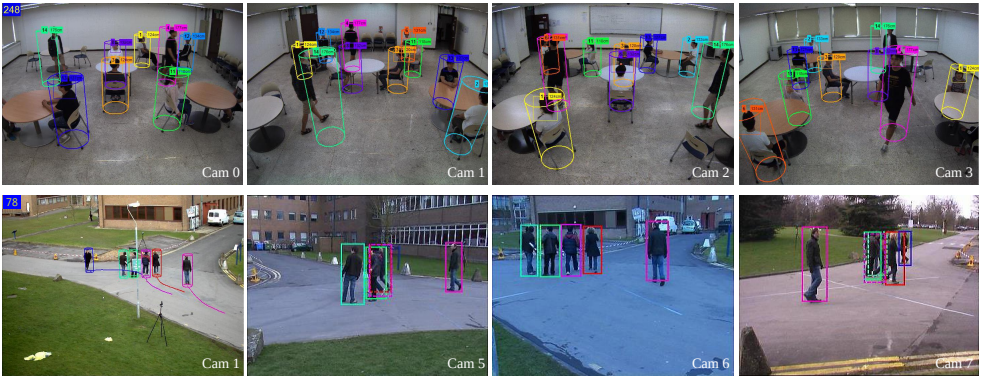

Figure 2: Qualitative results on *standing & sitting* sequence of the PSN-University dataset and PETS2009. 3D cylinder model of each person is projected to each camera. Our method can estimate the height of each individual as shown in *standing & sitting* sequence.

16 GB RAM in MATLAB, our method took 7.3 s/frame on an average until our algorithm converges, which is 8 times faster than 59 s/frame in [11]. For BIP based approaches [11, 13], the complexity grows exponentially with the number of cameras since combinations of observations from different cameras exponentially increase and their costs need to be predefined for a BIP solver. In our algorithm, by computing local changes of the costs, the cost calculation and the use of memory are much more efficient than the BIP based approaches [11, 13].

5 Conclusion

We have proposed an efficient approximation algorithm for MDA problem to solve spatio-temporal data association problem for multi-camera multi-target tracking. The approximation algorithm of MDA problem iteratively improves a feasible solution by two operations: random splitting and optimal merging. To improve the performance and reduce the computation, we defined a new cost function considering 3D reconstruction accuracy, motion smoothness, visibility from cameras, starting/ending at entrance and exit zone, and false positive. In particular, the proposed high-order motion model and 3D trajectory construction with 3D cylinder model can reduce the possibility of ID switches. As shown in the experiments, the proposed approximation method shows state-of-the-arts performance with 8 times faster computation than the existing BIP approach.

Acknowledgment

This work was supported by the IT R&D program of MOTIE/KEIT. [10041610, The development of automatic user information (identification, behavior, location) extraction and recognition technology based on perception sensor network (PSN) under real environment for intelligent robot] and the Brain Korea 21 Plus Project in 2015.

References

[1] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple Object Tracking Using K-Shortest Paths Optimization. *IEEE Trans. PAMI*, 33(9):1806–1819, 2011.

[2] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10, 2008.

[3] E. Brau, J. Guan, K. Simek, L. Del Pero, C. R. Dawson, and K. Barnard. Bayesian 3D Tracking from Monocular Video. In *ICCV*, 2013.

[4] M. Bredereck, X. Jiang, M. Korner, and J. Denzler. Data association for multi-object Tracking-by-Detection in multi-camera networks. In *ICDSC*, 2012.

[5] R. T. Collins. Multitarget data association with higher-order motion models. In *CVPR*, 2012.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom. A generalized S-D assignment algorithm for multisensor-multitarget state estimation. *IEEE Trans. Aerospace and Electronic Systems*, 33(2):523–538, April 1997.

[8] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *IEEE Trans. PAMI*, 36(8):1532–1545, August 2014.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010.

[10] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera People Tracking with a Probabilistic Occupancy Map. *IEEE Trans. PAMI*, 30(2):267–282, 2008.

[11] M. Hofmann, D. Wolf, and G. Rigoll. Hypergraphs for Joint Multi-view Reconstruction and Multi-object Tracking. In *CVPR*, 2013.

[12] S. M. Khan and M. Shah. Tracking Multiple Occluding People by Localizing on Multiple Scene Planes. *IEEE Trans. PAMI*, 31(3):505–519, 2009.

[13] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Branch-and-price global optimization for multi-view multi-target tracking. In *CVPR*, 2012.

[14] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *ICPR*, 2008.

[15] Y. Li, C. Huang, and R. Nevatia. Learning to associate: HybridBoosted multi-target tracker for crowded scene. In *CVPRW*, 2009.

[16] A. Milan, S. Roth, and K. Schindler. Continuous Energy Minimization for Multitarget Tracking. *IEEE Trans. PAMI*, 36(1):58–72, 2014.

[17] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38, 1957.

[18] A. B. Poore. Multidimensional assignment formulation of data association problems arising from multitarget and multisensor tracking. *Computational Optimization and Applications*, 3(1):27–57, March 1994.

[19] H. Possegger, S. Sternig, T. Mauthner, P. M. Roth, and H. Bischof. Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities. In *CVPR*, 2013.

[20] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof. Occlusion Geodesics for Online Multi-object Tracking. In *CVPR*, 2014.

[21] Z. Wu, N. I. Hristov, T. L. Hedrick, T. H. Kunz, and M. Betke. Tracking a large number of objects from multiple views. In *ICCV*.

[22] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.