2013 IEEE International Conference on Robotics and Automation (ICRA)
Karlsruhe, Germany, May 6-10, 2013

Global Optimal Data Association for Multiple People Tracking

Lili Chen[1], Wei Wang[2] and Alois Knoll[1]

Abstract— **Multiple people tracking is an important component for different tasks such as video surveillance and human-robot interaction. In this paper, a global optimization approach is proposed for long-term tracking of an *a priori* unknown number of targets, particularly aim to improve the robustness in case of complex interaction and mutual occlusion. With a state-space discretization scheme, the multiple object tracking problem is formulated with a grid-based network flow model, resulting in a convex problem that can be casted into an Integer Linear Programming (ILP), then solved through relaxation. In order to allow recovery from misdetections, common heuristics such as non-maxima suppression is eschewed within observations. In addition, we show that how behavior cue can be integrated into the association affinity model, providing discriminative hints for resolving ambiguities between crossing trajectories. The validity of the proposed method is demonstrated through experiments on multiple challenging video sequences, using a calibrated multi-camera setup.**

I. INTRODUCTION

People tracking is an important issue in various applications, such as video surveillance and cognitive human-robot interaction. However, it is a highly challenging problem, due to the uncertainty in physical target appearance, complex target interaction, mutual occlusion, cluttered environment. Tracking-by-detection approaches, with the advantage of being resistant to divergence, have demonstrated impressive results in addressing these challenges. Such approaches involve two separate steps, including time-independent detection and association of detection across frames.

The data association component is difficult in the face of false positives, missing detections, similar and mutually occluded targets. Classic data association approaches such as Global Nearest Neighbor (GNN) [1] is based on the idea of bipartite matching, which formulates the single-scan observation-to-track association as a two-dimensional assignment, choosing the one with the highest joint probability as final association for current scan among all possible assignments. It has low computational complexity, however, it suffers from severe drawbacks in dense and noisy environments. Other approaches, such as Joint Probabilistic Data Association Filters (JPDAFs) [2] and Multi-Hypothesis Tracking (MHT) [3] jointly consider the data association from sensor measurements to multiple overlapping tracks. In particular, JPDAF combines all of the potential measurements into one weighted average, before associating it to the track, in a single update. While MHT calculates every possible update hypothesis, with a track, formed by previous hypotheses associated to the target. Both methods are known to be quite complex, and require a careful implementation in terms of parameters. In particular, the latter can not avoid the drawback of an exponentially growing computational complexity, with the number of targets and measurements involved in the resolution situation. Moreover, a global optimal solution cannot be guaranteed in sub-exponential time although they attempt to model the joint trajectories of all objects.

Recent works show that global optimization approaches of using Dynamic and Linear Programming have appeared to be powerful alternatives. Berclaz et al. [4] studies an efficient approximate dynamic programming scheme over invidual trajectories. Greedy strategies are utilized to combine trajectories and handle potential conflicts. This approach tends to mix trajectories when targets are densely located, as occlusions are not explicitly modeled because of separate optimization.

By contrast, Linear Programming seeks to optimize all trajectories simultaneously over the whole sequence. Jiang et al. [5] tackles multiple people tracking problem with the use of Integer Linear Programming, in which the problem is formulated as multi-path searching by explicitly modeling the track interaction and objects' mutual occlusion. The metric for inter-object interaction term is convex while the intra-object term quantifying object state continuity through sequence. This scheme explores a large search space efficiently and gives a near-global optimality, because of the specific structure of the formulation. However, its state-space only consists of observations, not able to interpolate trajectories smoothly in case of the false alarms, moreover, it requires *a priori* knowledge of the number of targets, which severely limits its applicability in pratical situations.

Similarly, Berclaz et al. [10] formulates multi-people tracking problem as a constrained flow optimization, resulting in a convex problem that can be solved by standard Linear Programming techniques. Their method does not need *a priori* knowledge of target numbers, and the model is far simpler. Nevertheless, they haven't incorporate appearance features into data association process, which makes their approach prone to ID-switches in complicated scenarios. While dynamic model is also discarded in this work.

Shitrit et al. [11] extends the work of [10], that addresses the appearance limitation by exploiting the global appearance constraints, in which the total number of tracked person is partitioned into L groups, and a separate appearance is assigned to each group. It reduces the number of ID switches

[1]L. Chen and A. Knoll are with Department of Informatics, Technische Universität München, 85748 Garching bei München, Germany. {chenlil, knoll}@in.tum.de

[2]W. Wang is with the Faculty of Electrical Engineering and Information Technology, Technische Universität München, 85290 München, Germany. wangwei@lsr.ei.tum.de

(a) (b)

Fig. 1. Sample frames of close interaction and highly occlusion.

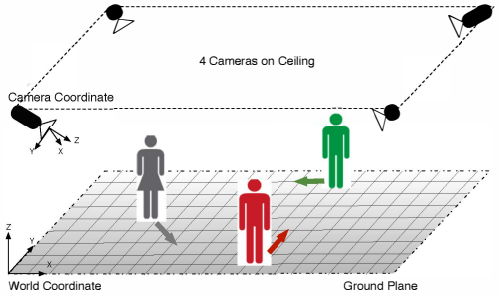


Fig. 2. Hardware setup.

for overlapping tracks. However, the appearance templates are selected manually through bounding boxes corresponding to members of each group.

Some other methods, like Quadratic Boolean Programming (QBP) [12], min-cost flow [6], have also been tailored to simultaneously optimize all tracks in polynomial time, are in fact closely related to ILP. The work [12] couples detection and estimation of trajectory hypotheses by QBP, such approach can only optimize over a limited time window, as the hypotheses search space is combinatorial. While Zhang et al. [6] defines data association as a maximum-a-posteriori (MAP) problem, and models trajectory hypotheses as disjoint flow paths in a cost-flow network.

Despite of intensive studies, robust and efficient tracking of multiple targets with complex interactions and significant mutual occlusions remains a problem. Meanwhile, the proposed different ways of handling the data association problem do not take advantage of any behavior cue, such as body orientation, which provides the direct evidence of what the person is going to do and where the person is facing at. In particular, it can provide valuable insight into the dynamics in case of social interaction and mutual occlusion. Although some works couple the dynamic model into the affinity model [13], however, such dynamic models mostly suppose steady heading, steady velocity or steady acceleration. It is problematic when a person is static or in low speed, as the velocity becomes too noisy to provide reliable information.

In this paper, we propose a global optimization approach for long-term tracking of an *a priori* unknown number of targets, with random walking in an overlapping, multi-camera environment. The primary goal is to address the problem of complex interaction and mutual occlusion by exploiting a consistency scheme on behavior cue, as well as compensating measurements of location and appearance. Fig. 1 shows the example of such difficulties that targets are interacted extremely close or highly occluded by each other, with similar position even appearance, and remains almost static over a few frames, nevertheless, the behavior cue - body orientation provides discriminative hints with corresponding targets.

More precisely, the multiple target tracking problem in this work is formulated in terms of finding the global maximum of a convex objective function, then solved efficiently through a linear programming relaxation. By taking the full advantage of our previous hierarchical grid-based detector [8], the regular discretization scheme is further adapted in current work. With this particular scheme, a grid-based network flow model is constructed, in which the nodes and edges encoded correspondingly, as also inspired by the work of [10]. This scheme allows to effectively avoid intermediate hard decisions and simply model mutual occlusion because of the specific graph structure. To enable the tracker recovering from mis-detections, we carry out non-maxima suppression during tracking rather than during detection, with the contrast to previous approaches that the state-space only consisting of observations, which are not able to interpolate trajectories smoothly in case of false negatives. Moreover, the measurements of body orientation, target location and appearance are incorporated in a global manner. The explicit use of behavior cue can disambiguate the situation such as in Fig. 1. This is distinctive compared to many state-of-the-art approaches that only depend on appearance or dynamic model.

The remainder of the paper is organized as follows. Section II describes the general system overview with hardware setup and algorithmic flow of software. The problem formulation and optimization framework are given in Section III. Section IV presents and discusses experimental results. At last, in Section V the paper is brought to conclusion and future development roads are proposed.

II. SYSTEM OVERVIEW

The system for the global optimal multiple people tracking is described, starting from the hardware setup to an overview of the proposed approach, followed by details on the specific components that are involved.

Our system hardware setup is depicted in Fig. 2. Four uEye usb cameras, with a resolution of 752×480, are mounted overhead on the corners of the ceiling, each of them observing the same 3D scene synchronously from different viewpoints, providing a more informative measurement set. Furthermore, all the four cameras are connected to one multi-core PC. A necessary step before being able to get accurate 3D information, is calibration of the intrinsic and extrinsic camera parameters, that we perform with the Matlab Calibration Toolbox, with respect to a *world* coordinate system placed on the floor.

The flow chart of our proposed approach is outlined in Fig. 3. After acquisition of original frames from the four

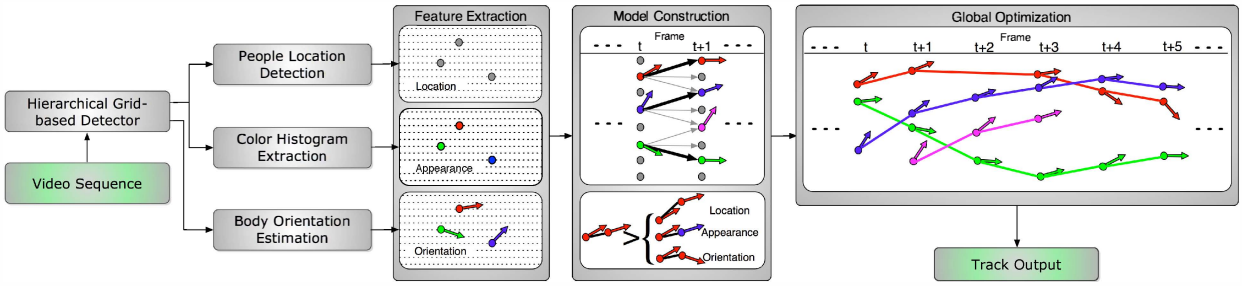


Fig. 3. Overview of the proposed approach.

cameras, a hierarchical grid-based detection [8] is followed, to obtain the potential observations. Note that we eschew common heuristics such as non-maxima suppression during detection, in order to allow following tracker to recover the most probable locations in accordance with all evidences. With the output from detector, each observation is characterized by a descriptor that records the features including location and appearance. However, it is not sufficient for a people tracking approach to determine data association only according to the location reference and appearance model, e.g. tracking may fail if two targets get very close or wear similar clothing. To overcome this limitation, we incorporate a discriminative cue on body orientation, which is estimated by utilizing the technique proposed in the work [9]. As proposed in our previous work [8], the state space is partitioned into integral grids with a coarse-to-fine strategy. We follow the discretization structure in current work, with the per-frame measurements sampled on regular grids. A grid-based network model can be constructed afterwards as concisely illustrated in model construction part, while the corresponding detailed model will be shown in Subsection III-A. A consistency scheme on behavior cue, as well as measurements of location and appearance, is modeled as transitional cost between nodes at two consecutive time steps. As illustrated in model construction part, the affinity measure can achieve highest only if the nodes have simultaneous similarity on all cues of location, appearance and orientation. Next follows the global optimization part, consists of formulating the data association problem as finding the global maximum of a convex objective function, which in our work is solved by a linear programming relaxation, and at last leading to track output with identity associated to each target.

III. GLOBAL OPTIMAL DATA ASSOCIATION

In this section, more details are provided about the proposed approach on finding global optimal solution for multi-target tracking. We start with the formulation of a grid-based network flow model, with the nodes and edges encoded. Then transform the maximum a-posteriori trajectory estimation into an Integer Linear Programming (ILP) problem, solved through relaxation. Followed by the association affinity model, in which a consistency scheme is exploited on behavior cue, as well as the compensation with measurements of location and appearance.

A. Grid-based Network Model

The state space is partitioned into discrete regions with a coarse-to-fine strategy during detection phase [8]. Each discrete region $\{R^{i,l}\}_{i=1}^{N_l}$ is sampled at its center, where $1 \le l \le L$, L is total levels of state space hierarchy, N_l is the number of grids at level l. With the refinement through detection, a set of observations with world-space position then would be on the leaf level L. Assume there are N_o^t observations at time instant t, $1 \le t \le T$, the observation set then be $R(t) = \left\{(R_t^{1,L}, R_t^{2,L}, \cdots, R_t^{N_o^t,L}\right\}$, while the full set of observations is $\Re = \{R(t)\}$. As we avoid non-maxima suppression during detection phase, these observations may contains many false positives, therefore, we wish to find a track for each target by eliminating the false positives and recovering from false negatives, as an ordered set of observations $T_n = \left\{(R_1^{n_1,L}, R_2^{n_2,L}, \cdots, R_T^{n_k,L}\right\}$, where $R_t^{n_i,L} \in \Re$, and the set of all single trajectories is, $\mathscr{T} = \{T_n\}$.

Due to the similar discretization strategy, an idea to construct a grid-based flow model is inspired by the work [10], with extension of a new consistency scheme within time intervals. For N_L discrete grids and T consecutive time steps, a directed acyclic graph (DAG) with $N_L T$ nodes is introduced as shown in Fig. 4, in which every node represents a discrete grid at a given time step. For a simpler flow-based analysis, the nodes are represented in the form of pairs within our model, allowing to explicitly model object dynamics through transition costs by considering the relationship of observations between two consecutive time steps. Whereas the transition cost in the model of [10], is assigned only with the occupancy probability of corresponding grid. For any location $R^{i,L}$, that an object located at $R^{i,L}$ (which will be encoded as node i in following text) at time t can reach its neighbors $\mathscr{N}(i)$ including itself at time $t+1$. Therefore, a path for the object starting from node i to node j is represented as $p_t^{i,j}$, valued $p_t^{i,j} \in \{0,1\}$, encoding that if the path is within part of some trajectory, that is, $p_t^{i,j} = 1$ means that the path is on the trajectory, and $p_t^{i,j} = 0$ means not. The cost $c(i,j)$ of each $p_t^{i,j}$ between node i and node j is assigned in the light of an association affinity model, which will be further described in Subsection III-C.

By taking advantage of the grid-based network flow model, we define a list of constraints to guarantee that each edge(path) through the DAG is practically possible:

Continuity Potential As illustrated in Fig. 5, in order to

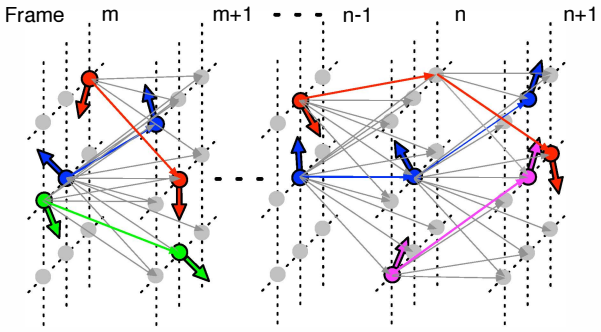


Fig. 4. The grid-based network flow model for multiple object tracking.

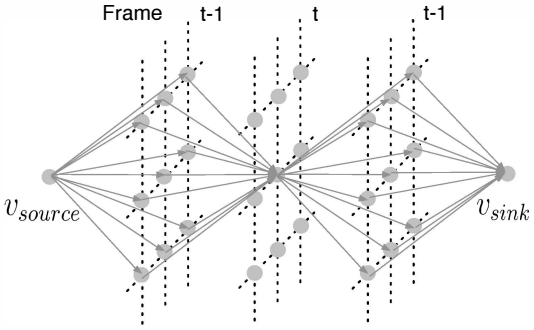


Fig. 5. Illustration of constraints.

enforce continuous trajectories for tracks, that for any node j, paths arriving at j at time t should be equal to the sum of paths leaving from j at time $t+1$,

$$\forall t, j, \sum_{i:j\in\mathcal{N}(i)} p_t^{i,j} = \sum_{k\in\mathcal{N}(j)} p_{t+1}^{j,k} . \tag{1}$$

Occlusion Term With the sampled grid resolution is sufficiently fine, no two objects should occupy the same grid at one time, thus, for any node j, the sum of paths from j should be no more than 1,

$$\forall t, j, \sum_{k\in\mathcal{N}(j)} p_t^{j,k} \le 1 . \tag{2}$$

Initialization and Termination Scheme For automatically initialize and terminate a track, a source and a sink nodes – v_{source} and v_{sink}, are introduced into the proposed network flow model, as shown in Fig. 5. At the first frame each node is connected to the source node, while at the last frame each one is connected to the sink. The source and sink nodes are subject to a contraint that all paths should start from v_{source} and end at v_{sink},

$$\sum_{j\in\mathcal{N}(v_{source})} p^{v_{source},j} = \sum_{k:v_{sink}\in\mathcal{N}(k)} p^{k,v_{sink}} . \tag{3}$$

B. Linear Programming Formulation

The objective of global optimal tracking is to link all the detections together over the whole sequence, choosing links so that the total probability is maximized, that is, maximizing the posteriori probability of $\mathcal{T}$ with given observation set $\Re$,

$$\begin{aligned} \mathcal{T}^* &= \underset{\mathcal{T}}{\operatorname{argmax}} P(\mathcal{T}|\Re) \\ &= \underset{\mathcal{T}}{\operatorname{argmax}} P(\Re|\mathcal{T})P(\mathcal{T}) . \end{aligned} \tag{4}$$

To convert it to an Integer Linear Programming(ILP) problem, its objective function is linearized with respect to a set of flows $p_t^{i,j} \in \{0,1\}$, which indicate if a path is within part of some trajectory or not, as mentioned earlier. Then the proposed grid-based network flow model can be expressed as an ILP with the following objective function, by minimizing the total cost,

$$\begin{aligned} X^* &= \underset{X}{\operatorname{argmin}} C^T p \\ &= \underset{X}{\operatorname{argmin}} \sum_i c(v_{source}, i) p^{v_{source},i} + \sum_{i,j,t} c(i,j) p_t^{i,j} \\ &\quad + \sum_i c(i, v_{sink}) p^{i,v_{sink}} , \end{aligned} \tag{5}$$

in which the cost function C will be described in more details in Subsection III-C.

Minimizing the criterion of (5) under the constraints of (1) to (3) can be rewrote as follows,

$$\begin{aligned} \text{minimize} \quad & C^T p \\ \text{subject to} \quad & \forall t, j, \sum_{i:j\in\mathcal{N}(i)} p_t^{i,j} = \sum_{k\in\mathcal{N}(j)} p_{t+1}^{j,k} \\ & \forall t, j, \sum_{k\in\mathcal{N}(j)} p_t^{j,k} \le 1 \\ & \sum_{j\in\mathcal{N}(v_{source})} p^{v_{source},j} = \sum_{k:v_{sink}\in\mathcal{N}(k)} p^{k,v_{sink}} \\ & \forall t, i, j, \; p_t^{i,j} \in \{0,1\} . \end{aligned} \tag{6}$$

Since Integer Linear Programming is NP-complete, we relax the condition $p_t^{i,j} \in \{0,1\}$ to $0 \le p_t^{i,j} \le 1$, resulting in a significant complexity reduction, and the relaxed formulation can be sufficiently solved with the simplex or interior-point method. The LP results then, are no longer guaranteed to be integer. However, we find in the experiments that the results are in most cases round integral, therefore gives the globally optimized solution.

C. Association Affinity Model

The details on the association affinity model are provided, which incorporate the measurements on behavior cue, as well as location and appearance in a global manner. With the set of observations $\Re$, we extract the features respect to location, color appearance, human body orientation, as illustrated in feature extraction module in Fig. 4. Location feature of each observation is represented as a grey node, its corresponding color measurement is represented as a colored one, while the arrow indicates the orientation cue.

Therefore, the transition probability term for each path $p^{i,j}$ leaving from node i to node j, is according to,

$$A(i,j) = \begin{cases} \lambda_1 \cdot A_{pos}(i,j) + \lambda_2 \cdot A_{appr}(i,j) + \lambda_3 \cdot A_{ori}(i,j), \\ \qquad\qquad if\ t_j - t_i = 1, j \in \mathcal{N}(i) \\ 0, \qquad\qquad otherwise \end{cases}, \tag{7}$$

which is a weighted sum of these three affinities $A_{pos}(i,j)$, $A_{appr}(i,j)$, $A_{ori}(i,j)$, respectively are location, appearance, orientation affinity between nodes i and j. λ_1, λ_2 and λ_3 are constant coefficients, with $\lambda_1 + \lambda_2 + \lambda_3 = 1$, to control the weight for compensating with each other if any of the three features becomes ambiguous. To minimize the total cost according to (6), $-log(A(i,j))$ is defined as the cost $c(i,j)$ of path $p^{i,j}$.

In particular, the location affinity term $A_{pos}(i,j)$ concerns the spatial distances between two detection responses within two consecutive time steps,

$$A_{pos}(i,j) = exp(-\frac{||l_i - l_j||}{\sigma_l^2}) . \tag{8}$$

Note that the detection responses are on the 3D ground plane, not in 2D image plane. And the absolute spatial location difference is a L_1 norm.

For the appearance term, CIE Lab color space is employed for better characterizing the color content, which has the advantage of being perceptually uniform. $64 \times 64 \times 64$ color histograms are extracted from foreground images according to the detection responses. It is worth noting that the foreground images are obtained through utilizing a GPU based foreground/background segmentation approach proposed by Griesser et al. [14].

To compare the color feature similarity, Bhattacharyya distance measure is utilized because of its good classification property, allowing the combination of different features in a straightforward way. The similarity is multiplied through all views and assigned to corresponding path between nodes i and j.

$$A_{appr}(i,j) = \prod_{n_v} exp(-\frac{d_B(a_i,a_j)}{\sigma_a^2}) , \tag{9}$$

where d_B is the Bhattacharyya distance between color feature a_i and a_j.

Follows the crucial body orientation term, which is seldom considered in most state-of-the-art approaches. As already emphasized before, the body orientation cue provides hints for resolving ambiguities between crossing trajectories, which is discriminative enough even if crossed targets have similar appearance or move very slow.

This cue can be estimated with location reference from detection responses. Therefore, the affinity is defined by the difference between two consecutive orientation estimates,

$$A_{ori}(i,j) = exp(-\frac{0.5 * (1 - \cos(|\theta_i - \theta_j|)}{\sigma_\theta^2}) . \tag{10}$$

Note that the orientation θ_i and θ_j are computed in 3D space, being defined as the rotation with the axis perpendicular to the ground plane. The form of $0.5*(1-\cos(|\theta_i - \theta_j|))$ makes the orientation difference lie in the interval of $[0,1]$.

The significant advantage of adding the orientation affinity term, is that more accurate trajectories can be estimated in case of close interaction or mutual occlusion, which will be demonstrated in our experiments in Section IV.

IV. EXPERIMENTAL RESULTS

This section aims to show the demonstrative results of our proposed approach. We evaluate the algorithm through pre-recorded video sequences, with multiple people entering and leaving the scene, as well as closely interacting with each other for long time, or be seriously occluded by others. The sequences have been simultaneously recorded from four cameras, as described in Section II, with a resolution of (752×480), and a frame rate of 25 fps.

A. Implementation Details

Before carrying out the global optimization for data association, the Hierarchical Grid-based Detection algorithm [8] and 3D Appearance Model based Body Orientation Estimation [9] are performed, to obtain the required observations and features. For completeness of the paper, we briefly summarize the steps below.

Hierarchical Grid-based Detection It first performs an edge-based background subtraction with images from all views. Afterwards, an oriented distance transform is computed on foreground edge image, in order to match with each template, through both location and orientation of its contours' each pixel. The state-space is partitioned into discrete regions with a coarse-to-fine strategy, the templates are then generated by rendering a 3D model composed of 3 cylinders at each state, under respective camera projection. The likelihoods are then computed by matching projected templates and oriented DT for each camera view on coarse grid firstly, then refined on the next resolution only the locations where its likelihood is higher than a given threshold, the joint likelihoods can simply be multiplied over all views. Therefore, at each time t and each discrete region (x,y), the likelihood $\rho_t^{x,y}$ is produced as the probability of presence of a target.

In our experiments, the state grids are set up respectively as 10×10, 20×20 and 40×40 from the coarsest to the finest, resulting in a total of 2100 grid cells. Since the area of interest is $(6m \times 4.2m)$, the corresponding grid on the finest level has a resolution of $(150mm \times 105mm)$.

Body Orientation Estimation With the ground plane location observations from above grid-based detector, the body orientation θ can be estimated through the method proposed in [9] with the location reference (x,y). In a nutshell, it generates a 3D appearance model for each new target, by back-projecting the pixels from foreground images onto the surface of 3D geometry body model, rendered as a 3D colored point cloud, then combining with a 2D template-based matching approach due to 3D/2D projection and visibility test.

Orientation θ is defined as the rotation of the minor axis of cross section of 3D body model with the axis perpendicular

to the ground plane, while in our experiments, it is discretized into 12 discrete orientations by covering 360°.

Global Optimal Data Association As described above, the space is discretized into 40×40 grids on the finest level, with each node of the grid at time t connecting to its 9-neighborhood (8 neighbors and the central location itself) at time $t+1$, resulting in 14,400 flows between two consecutive frames. We define the transition cost $c(i, j)$ to be 0 if there is no observation on node i, that reduces the size of graph, which efficiently decreases the computational cost. The three parameters σ_l, σ_a and σ_θ within association affinity model are the standard deviation, all set to 0.5 empirically, governing the relative influence of the similarity corresponding to location, color and orientation. Moreover, in order to make the optimization process tractable for long sequences, we utilize the common strategy of separating the sequence into several batches of frames, with an overlapping time window. The number of frames in each batch and the overlap length are respectively set to 50 and 10. And the LP problem is solved by IBM ILOG CPLEX Optimizer [15].

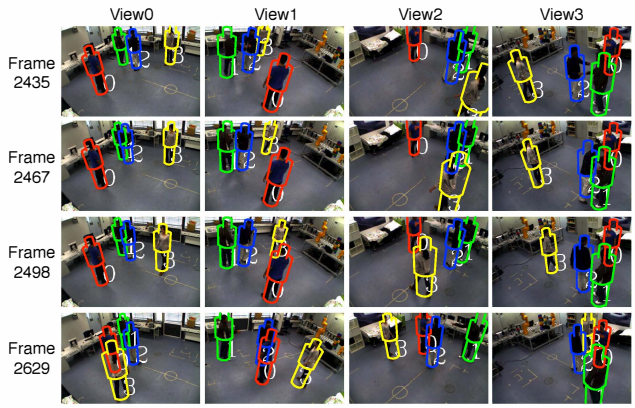
B. Tracking Performance

Two sets of experiments are conducted, by testing on two sequences both with four targets involved, while the observing area is $6m \times 4.2m$.

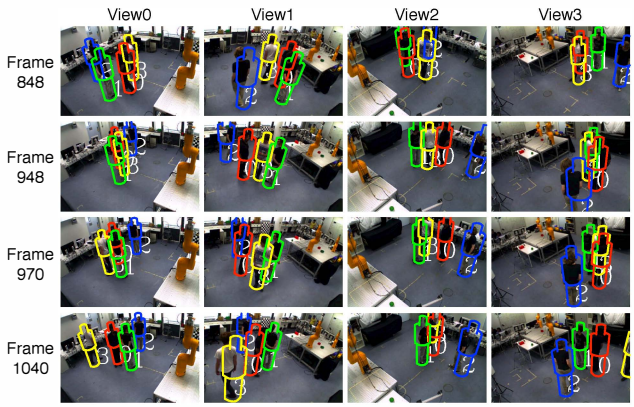
The first experiment tests with a sequence consists of 3160 frames, in which the objects have interaction for long time. This scenario is aiming at evaluating the ability of our approach for dealing with long-term interaction, especially to verify the validity of the affinity term on behavior cue. Fig. 6(a) illustrates some sample frames of the tracking result, particularly between frame 2435 and frame 2498, target 1 and target 2 get extremely close and interact with each other across several frames, even with their clothing quite similar. Nevertheless, as they have obvious opposite body orientation, which can provide a powerful compensation in this ambiguous case, target 1 and 2 successfully maintain their own identity throughout the interaction.

The second set of experiments is conducted on a sequence with 1800 frames. Within this sequence, most of the targets are wearing very dark clothing, with ambiguous appearance compared to each other. The objective of this case is to evaluate the capability of how the three affinity terms compensate with each other if any of the three features becomes ambiguous. As shown in Fig. 6(b), we can see the challenges due to targets that are occluded by each other from one or two views. Between frame 948 and frame 970, target 3 is trying to pass through between target 0 and 1, their spatial locations are very close, however the distinguishing appearance of target 3 helps itself maintain its identity, as well as owing to its different orientation compared to other two targets. Conversely at frame 1040, target 0 and target 1 have close interaction while wearing extreme similar dark clothing, however their distinctive orientation provides efficient hint despite of the similar appearance.

Note that there are approximately 100 observations in each frame by eschewing non-maxima suppression during



(a) Sequence 1



(b) Sequence 2

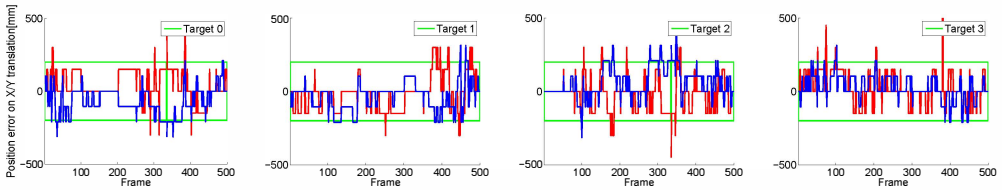
Fig. 6. Tracking results of our proposed approach on four camera views.

detection phase. The adequate observations also greatly help for preserving the tracks during heavy occlusion and long-term interaction.

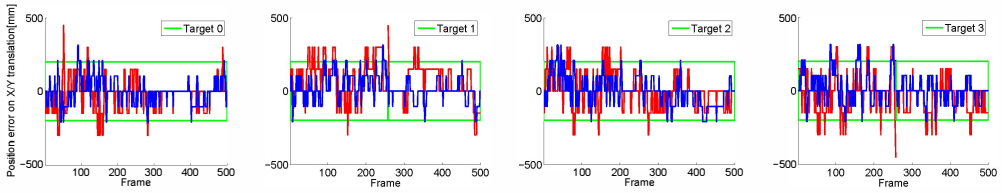
During conducting the experiments, we also pay special attention on the result values of the variables in linear programming relaxation. By processing each batch of 50 frames, which results in 720,000 variables, we notice that 719,800 of which are in the range $[0, 0.01]$, while 200 in the range $[0.99, 1]$. That means the relaxed linear programming is able to give a global optimal solution in real problems.

C. Quantitative Evaluation

To better evaluate the performance of our proposed approach, we manually label ground truth data for each frame of the sequences, by rendering 3D cylinder model to coincide with the target area within image. Fig. 7 gives a quantitative evaluation of our experimental results, with the terms of identity maintainance and position accuracy. We select the most challenging clip including 500 frames for each sequence. (X, Y) position errors are illustrated in red and blue lines respectively, while the green boxes indicate sub-tracks. As expected, our global optimal data association based on multi-featured affinity model greatly improves the tracking performance with significant sub-track reduction, nicely filters the grid-based detection results, smoothly linking detections



(a) Sequence 1



(b) Sequence 2

Fig. 7. Performance evaluation with ground truth data.

TABLE I
MEAN SQUARED POSITIONING ERROR

	Target 0	Target 1	Target 2	Target 3
Sequence 1	112.86 mm	101.74 mm	107.55 mm	144.58 mm
Sequence 2	76.39 mm	123.94 mm	105.19 mm	101.69 mm

together in case of mis-detections and false positives. In sequence 1, target 0 and 2 smoothly switches id for one time around frame 2635, due to aforementioned occlusions and interactions, target 1 and 3 successfully maintains its identity throughout 500 frames. Similarly in sequence 2, target 1 and 3 switches their id around frame 838, while target 0 and 2 keeps well its identity.

The position accuracy from Fig. 7 indicates that, despite the cluttered situation, the position errors are considerably low for each target, being most of the time under $100-150mm$. To be more quantitative, we also give out the mean squared positioning error for each target within corresponding sequence in Table I. As mentioned above, the resolution on the finest grid is $(150mm \times 105mm)$, therefore the error corresponds to approximately one grid.

For the running time, the optimization process for each batch of 50 frames takes 6.3s while executed on a desktop PC with Intel Core 2 Duo CPU (1.86GHz) and 3GB memory.

V. CONCLUSIONS

In this paper, we have proposed a global optimization framework for tracking a varying number of targets on discrete grids. Multiple target tracking problem is casted into Integer Linear Programming and then solved through relaxation, achieving a global-optimality in most cases. Experimental results on two challenging video sequences, demonstrate that our proposed approach deals fairly well with mutual occlusions and long-term interactions. The ground truth data is annotated for better performance evaluation. The proposed methodology can easily be applied to different camera setup and different environment, while also additional features can be included. Future work may involve investigating the optimization scheme further, with improving speed, robustness and versatility. In addition, we would like to extend our framework with non-stable illumination conditions, and also apply the system to high-level scenarios, such as analysis of the trajectories, as well as human robot interaction applications.

REFERENCES

[1] F. Burgeois, An Extension of the Munkres Algorithm for the Assignment Problem to Rectangular Matrices, Communications of the ACM, vol. 14, pp. 802-806, 1971.
[2] T. E. Fortmann, Y. Bar-Shalom and M. Scheffe, Sonar Tracking of Multiple Targets using Joint Probabilistic Data Association, IEEE Journal of Oceanic Engineering, vol. 8, no. 3, pp. 173-184, 1983.
[3] D. B. Reid, An Algorithm for Tracking Multiple Targets, IEEE Transaction on Automatic Control, vol. 24, no. 6, pp. 843-854, 1979.
[4] J. Berclaz, F. Fleuret and P. Fua, Robust People Tracking with Global Trajectory Optimization, IEEE Conference on Computer Vision and Pattern Recognition, pp. 744-750, 2006.
[5] H. Jiang, S. Fels and J. J. Little, A Linear Programming Approach for Multiple Object Tracking, IEEE Conference on Computer Vision and Pattern Recognition, pp. 744-750, 2007.
[6] L. Zhang, Y. Li and R. Nevatia, Global Data Association for Multi-Object Tracking Using Network Flows, IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[7] Y. Li, C. Huang and R. Nevatia, Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene, IEEE Conference on Computer Vision and Pattern Recognition, pp. 2953-2960, 2009.
[8] L. Chen, G. Panin and A. Knoll, Multi-camera People Tracking with Hierarchical Likelihood Grids, Proceedings of the 6th International Conference on Computer Vision Theory and Applications, pp. 474-483, 2011.
[9] L. Chen, G. Panin and A. Knoll, Human Body Orientation Estimation in Multiview Scenarios, Proceedings of the 8th International Symposium on Visual Computing, 2012.
[10] J. Berclaz, F. Fleuret, P. Fua, Multiple Object Tracking Using Flow Linear Programming, Winter-PETS, 2009.
[11] H. B. Shitrit, J. Berclaz, F. Fleuret and P. Fua, Tracking Multiple People under Global Appearance Constraints, ICCV, 2011.
[12] B. Leibe, K. Schindler and L. V. Gool, Coupled Detection and Trajectory Estimation for Multi-object Tracking, ICCV, 2007.
[13] A. Andriyenko and K. Schindler, Globally Optimal Multi-target Tracking on a Hexagonal Lattice, ECCV, 2010.
[14] A. Griesser, D. S. Roeck, A. Neubeck and L. Van Gool, Gpu-based Foreground-Background Segmentation using an Extended Colinearity Criterion, Proc. of Vision, Modeling, and Visualization(VMV), pp. 319-326, 2005.
[15] http://www.ibm.com/software/integration/optimization/cplex-optimizer.