Multiobject Tracking as Maximum Weight Independent Set

William Brendel, Mohamed Amer, Sinisa Todorovic
Oregon State University, Corvallis, OR 97331, USA
brendelw@onid.orst.edu, amerm@onid.oregonstate.edu, sinisa@eecs.oregonstate.edu

Abstract

This paper addresses the problem of simultaneous tracking of multiple targets in a video. We first apply object detectors to every video frame. Pairs of detection responses from every two consecutive frames are then used to build a graph of tracklets. The graph helps transitively link the best matching tracklets that do not violate hard and soft contextual constraints between the resulting tracks. We prove that this data association problem can be formulated as finding the maximum-weight independent set (MWIS) of the graph. We present a new, polynomial-time MWIS algorithm, and prove that it converges to an optimum. Similarity and contextual constraints between object detections, used for data association, are learned online from object appearance and motion properties. Long-term occlusions are addressed by iteratively repeating MWIS to hierarchically merge smaller tracks into longer ones. Our results demonstrate advantages of simultaneously accounting for soft and hard contextual constraints in multitarget tracking. We outperform the state of the art on the benchmark datasets.

1. Introduction

This paper addresses the problem of simultaneous tracking of multiple targets in a complex scene, captured by a non-static camera. Targets are occurrences of known object classes, such as cars, pedestrians, and bicycles. Every target is characterized by time-varying appearance and motion properties. Targets are also characterized by their spatiotemporal interactions, such as pedestrians moving in the same or opposite direction, and domain-specific constraints, such as pedestrians tend to move similarly but usually try to keep distance from one another. We refer to these interactions and constraints as context. Given a similarity (or distance) function in terms of target appearance, motion, and contextual properties, tracking can be formulated as matching similar object occurrences across video frames. Our goal is to:

1. Learn online the statistical intrinsic and contextual properties of objects to specify their similarity, and
2. Match similar object occurrences in consecutive frames by simultaneously accounting for their hard and soft contextual constraints.

We address a setting in which the number of targets, their class membership, and their layouts in the video may be arbitrary, and no training examples of these are available.

1.1. Relationships to Prior Work

Multitarget tracking is challenging, because the uncertainty about targets may arise from a multitude of sources, including: similarity of targets from the same class, complex target interactions, occlusions over relatively long time, and dynamic, cluttered backgrounds. Tracking-by-detection approaches have demonstrated impressive results in addressing these challenges [16, 8, 13, 22, 1, 10, 23, 3, 4]. They first apply an object detector to generate target hypotheses in each frame, and then transitively link the detections so as to maintain their unique identities. The transitive linking is difficult in the face of (potentially numerous) false positives and missing detections. This is usually addressed by learning an affinity model between detections in terms of their intrinsic properties (e.g., color, posture, speed, direction) [13, 1, 23, 11, 14], as well as spatiotemporal context [15], supporting evidence from neighboring tracks [9], and estimates of an occluder map [10] and 3D scene layout [4]. Given affinities between detections, the aforementioned work formulates tracking as the data association problem. This is typically posed as bipartite matching, with the constraint that the matching be one-to-one, and solved by either the greedy Hungarian algorithm, or more sophisticated network flow algorithms [24].

Beyond the one-to-one constraint, various relationships between objects give rise to other soft and hard constraints which can be used for tracking. This motivates us to extend prior work by incorporating additional contextual constraints in data association. We show that this extension naturally lends itself to the maximum weight independent set (MWIS) problem. For this more general formulation of multitarget tracking, we present a new MWIS algorithm.

Tracking-by-detection approaches may poorly perform in the presence of long-term occlusions, i.e., long gaps in a sequence of object detections. This can be addressed by fusing particle filtering with detector confidences for more accurate maintaining of tracking hypotheses [3]. Alternatively, the long gaps can be overcome by a hierarchical association of detections [10]. Brute-force strategies have been proposed to handle errors in the track linking by augmenting the initial set of tracks with their merges and splits [19].

We address the long gaps by iteratively linking smaller similar tracks into larger ones, and splitting long unviable tracks, while respecting their soft and hard contextual constraints, until convergence. Unlike [10], we conduct both merging and splitting of tracks, and thus allow corrections of any errors made in the previous iterations.

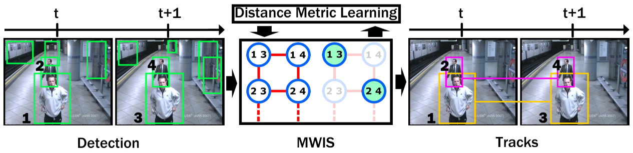


Figure 1. Our approach: Object detections are used to build a graph of detection pairs, called tracklets. Tracking is formulated as finding the maximum-weight independent set (MWIS) of the graph, and solved by our new MWIS algorithm. Similarity between detections and contextual constraints between the tracks are learned online. Long-term occlusions are addressed by iteratively applying MWIS to merge smaller tracks into longer ones.

1.2. Overview of Our Approach

Fig. 1 illustrates the following steps of our approach.

Step 1: We apply detectors of a set of object classes to all video frames. Each detection is characterized by a descriptor that records the following properties of the corresponding bounding box: location, size, and the histograms of color, intensity gradients, and optical flow.

Step 2: The best matching detections are transitively linked across video into distinct tracks, whose total number is unknown a priori. This is done under the hard constraint that no two tracks may share the same detection, to prevent implausible video interpretations. In addition, the linking is informed by spatiotemporal relationships between the tracks, which provide for soft constraints. To this end, we build a graph, where nodes represent candidate matches from every two consecutive frames, referred to as tracklets; node weights encode the similarity of the corresponding matches; and edges connect nodes whose corresponding tracklets violate the hard constraints. Given this attributed graph, data association is formulated as the maximum-weight independent set (MWIS) problem. MWIS is the heaviest subset of non-adjacent nodes of an attributed graph. Conveniently, MWIS of the entire graph is equivalent to a union of the MWIS solutions of independent subgraphs. This allows us to conduct multitarget tracking online. We present a new MWIS algorithm that is guaranteed to converge to an optimum.

Step 3: Intrinsic target properties and pairwise context, used in Step 2, are learned online, as the tracks keep accumulating statistical evidence of the targets. The relative significance of these properties for each track is learned so as to minimize the Mahalanobis distances of detections within the same track, and maximize the Mahalanobis distances between detections from distinct tracks.

Step 4: To address long-term occlusions, we iterate Step 2 and Step 3 to merge or split tracks so as to increase the total weight of the MWIS, until convergence.

1.3. Contributions

We formulate multitarget tracking as the MWIS problem. MWIS allows concurrent and direct reasoning about soft and hard contextual constraints, whereas prior work typically relaxes hard constraints to the continuous domain for tractability (e.g., [24]). Importantly, the MWIS formulation provides a principled way of partitioning the entire graph of candidate tracklets into independent subgraphs, which simplifies our data association problem to a number of smaller MWIS problems for each subgraph.

MWIS has also been used for tracking in [20], with many differences. They build a graph where each node represents an entire track hypothesis, whereas our nodes are tracklets. Our graph gets broken down into smaller independent subgraphs, which is not the case in [20]. They reformulate MWIS as a semi-definite program, and use a rank-constrained approximation to solve it, whereas we directly solve the exact MWIS formulation. Global optimal trajectory association has also been formulated as the min-cost flow problem in [24].

MWIS is a well-researched combinatorial optimization problem, known to be NP-hard, and hard to approximate. Numerous heuristic approaches exist. For example, iterated tabu search [17] uses a trial-and-error, greedy search in the space of possible solutions, with an optimistic complexity estimate of $O(n^3)$. MWIS is often reformulated as the maximum weight clique (MWC) problem that uses a dual graph of the original [18]. However, important hard constraints captured by edges of the original graph may be lost in this conversion.

We derive a new MWIS algorithm that iteratively refines the solution using a first-order dynamic. Also, we prove its convergence to a maximum, with complexity $O(n^2)$, where n is the number of nodes in a graph.

The remainder of the paper presents details of each step of our approach, starting from Step 1.

2. Object Detection

Given a video, we use a state-of-the-art detector to identify occurrences of target object classes in every frame.

We consider the following alternatives: (i) Implicit Shape Model (ISM) [12], (ii) HOG detector [5], and (iii) Deformable part-based model [7]. The same detectors have been used with success in prior work (e.g., [3, 4]).

Each detected bounding box, z, is characterized by a descriptor, $\boldsymbol{z}$, whose elements include: (a) location and size of the bounding box, and (b) a PCA projected vector at 5% reconstruction error of the following features: (b.i) HOG descriptor of size 81×1, (b.ii) HSV color histogram of size 256×3, and (b.iii) two 10-bin histograms of optical flow along x and y directions within the box.

Given two detections z and z', and their descriptors $\boldsymbol{z}$ and $\boldsymbol{z}'$, similarity between them is defined as:

$$w = \exp(-(\boldsymbol{z}-\boldsymbol{z}')^{\mathrm{T}}\mathbf{M}(\boldsymbol{z}-\boldsymbol{z}')), \tag{1}$$

where $\mathbf{M}$ is a distance metric matrix. $\mathbf{M}$ is initialized to the identity matrix, and then learned online (Sec. 4.1).

3. Data Association is the MWIS Problem

This section presents our Step 2. We first formalize data association, and then cast it as the MWIS problem. We also specify a new MWIS algorithm.

Let $Z^{(t)} = \{z_1^{(t)}, z_2^{(t)}, \ldots\}$ denote the set of object detections at time t, and $Z = \cup_{t=1,\ldots,T} Z^{(t)}$ be the set of all detections. A track is an ordered set of detections $\mathcal{T} = \{z_a^{(t_1)}, z_b^{(t_2)}, \ldots\}$, such that $\forall t, |\mathcal{T} \cap Z^{(t)}| \leq 1$.

Def. 1. *Data association is defined as the problem of finding a subset of all detections whose time sequences form a set of non-overlapping tracks, $\Sigma = \{\mathcal{T}_k : \mathcal{T}_k \cap \mathcal{T}_l = \emptyset,\ k \neq l,\ k, l = 1, 2, \ldots\}$, $\Sigma \subseteq Z$, such that each $\mathcal{T}_k \in \Sigma$ is a set of all detections of a unique target.*

The data association problem can be formalized by constructing a graph, $G = (V, E, w)$, illustrated in Fig. 2a. V is the set of nodes representing pairs of object detections from every two consecutive frames, called tracklets, $V = \{i^{(t)} : i^{(t)}{=}(z_a^{(t)}, z_b^{(t+1)}), z_a^{(t)}{\in}Z^{(t)}, z_b^{(t+1)}{\in}Z^{(t+1)}, t{=}1, \ldots, T\}$, with cardinality $|V|{=}n$. E is the set of undirected edges connecting only those tracklets $i^{(t)}{\in}V$ and $j^{(t)}{\in}V$ that happen at the same time $t{\to}t{+}1$, and share the same detection, $E = \{(i^{(t)}, j^{(t)}) : i^{(t)}{\cap}j^{(t)}{\neq}\emptyset,\ i^{(t)}{\neq}j^{(t)},\ t{=}1, \ldots, T\}$. Finally, $w : V{\to}\mathbb{R}^+$ associates positive weights w_i with every node $i \in V$, defined as similarity by Eq. (1). Note that tracklets from different time instances, e.g., $i{=}(z_a^{(t)}, z_b^{(t+1)})$ and $j{=}(z_b^{(t+1)}, z_c^{(t+2)})$, may share detection $z_b^{(t+1)}$, and still remain unconnected in the graph. Thus, by construction, G consists of a number of independent subgraphs, $G = \{G^{(t)} : t = 1, \ldots, T\}$.

Below, we prove that the data association problem is equivalent to finding the MWIS of G. It is easy to show that a track, $\mathcal{T}$, can equivalently be defined as an ordered set of tracklets, $\mathcal{T} = \{i^{(t_1)}, j^{(t_2)}, \ldots\}$, such that $\forall t, |\mathcal{T} \cap G^{(t)}| \leq 1$, and if two consecutive tracklets $i^{(t)}$ and $j^{(t+1)}$ belong to $\mathcal{T}$ then $i^{(t)}$ must end, and $j^{(t+1)}$ must start with the same detection (for maintaining track identity), as illustrated in Fig. 2b. In addition, it is straightforward to show that any two non-overlapping tracks $\mathcal{T}_k \cap \mathcal{T}_l = \emptyset$, can be formed only from independent tracklets, $\forall i \in \mathcal{T}_k, \forall j \in \mathcal{T}_l$, $(i, j) \neq E$. This allows us to state the following proposition.

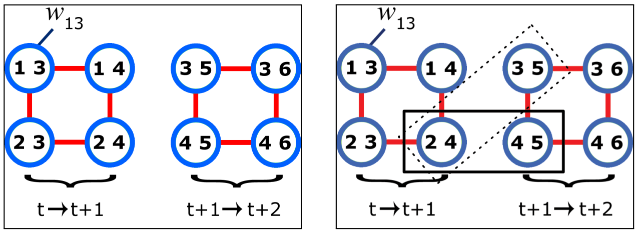


Figure 2. The graph: (a) Nodes in G are tracklets that are connected by edges if they happen at the same time $t{\to}t{+}1$, and share the same detection (denoted with integers); this partitions G into independent subgraphs. (b) A track (bold rectangle) consists of a time sequence of tracklets (the dashed track is forbidden).

Proposition 1. *The data association problem can be specified as finding a subset of all independent tracklets in G whose time sequences form a set of non-overlapping tracks, Σ, and whose total weight, $\sum_{i\in\Sigma} w_i$, is maximum.*

Proof. We use contraposition. Suppose that the MWIS of G, denoted as $\tilde{\Sigma}$, consists of tracklets whose time sequences do not satisfy Def. 1. By definition of independent set, the tracks in $\tilde{\Sigma}$ must be non-overlapping. Then, from Def. 1, there must exist a detection z in the video that does not belong to any track in $\tilde{\Sigma}$. By construction of G, it follows that there is a tracklet i that contains z, such that i is independent of all tracklets in $\tilde{\Sigma}$. Since tracklet weights are positive, $\tilde{\Sigma} \cup \{i\}$ is an independent set with larger total weight than $\tilde{\Sigma}$, which contradicts the initial assumption that $\tilde{\Sigma}$ is MWIS. □

Since the MWIS of G is equal to a union of the MWIS of each independent subgraph $G^{(t)}$, $t = 1, \ldots, T$, we first separately solve the MWIS of each $G^{(t)}$, denoted as $\Sigma^{(t)}$. Then, following the above definitions, we link tracklets into distinct tracks, such that a track, $\mathcal{T}$, may contain only one tracklet from each $\Sigma^{(t)}$, $t = 1, \ldots, T$, and $\mathcal{T}$ may contain two consecutive tracklets $i^{(t)} \in \Sigma^{(t)}$ and $j^{(t+1)} \in \Sigma^{(t+1)}$ only if $i^{(t)}$ ends and $j^{(t+1)}$ starts with the same object detection. In the following, we present a formulation of the MWIS problem, and specify a new MWIS algorithm.

3.1. The MWIS problem

A subset of V can be represented by an indicator vector $\boldsymbol{x} = (x_i) \in \{0, 1\}^n$, where $x_i = 1$ means that node i is in the subset, and $x_i = 0$ otherwise. Then, MWIS, denoted as

$\boldsymbol{x}^*$, is specified by the following integer program:

$$\begin{aligned}&\boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}}\, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x},\\ &\text{s.t. } \forall i \in V, x_i \in \{0,1\}, \text{ and } \forall (i,j) \in E, x_i \cdot x_j = 0,\end{aligned} \tag{2}$$

where $\boldsymbol{w} = (w_i)$ is the vector of node weights defined in (1). Note that instead of the quadratic constraints $\forall(i,j)$ $x_i \cdot x_j = 0$ in (2), one could use the linear constraints $\forall(i,j)$, $x_i + x_j \leq 1$. However, since (2) is typically solved by a relaxation to the continuous domain, the relaxed linear constraints would be much weaker than the quadratic ones. For example, with x_i=0.5 and x_j=0.5, we have 0.5+$0.5 \leq 1$, which still satisfies the linear constraint, whereas $0.5{\cdot}0.5 \neq 0$. The independence constraint in (2) can be directly incorporated in the objective function, which results in the following equivalent formulation:

$$\begin{aligned}&\boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}} \textstyle\sum_{i\in V} w_i x_i \prod_{j\in V,(i,j)\in E}(1 - x_j),\\ &\text{s.t. } \forall i \in V,\ x_i \in \{0,1\}.\end{aligned} \tag{3}$$

In (3), the sum does not increase for solutions in which both x_i and x_j are set to 1, and their corresponding nodes are connected in the graph, $(i,j) \in E$. The objective of (3) can be more conveniently written using the adjacency matrix of G, $\mathbf{B} = (\mathbf{B}_{ij})$, with elements $\mathbf{B}_{ij} = 1$ if $(i,j) \in E$, and $\mathbf{B}_{ij} = 0$ otherwise, as follows

$$\begin{aligned}&\boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}} \textstyle\sum_{i\in V} w_i x_i \prod_{j\in V}(1 - x_j)^{\mathbf{B}_{ij}},\\ &\text{s.t. } \forall i \in V,\ x_i \in \{0,1\}\end{aligned} \tag{4}$$

Eq. (4) gives the exact discrete formulation of the MWIS problem. As common in combinatorial optimization, we relax this discrete formulation to the continuous domain. Specifically, we introduce an auxiliary, real-valued vector, $\boldsymbol{y} = (y_i) \in \mathbb{R}^n$, and replace the constraint $\forall i{\in}V, x_i{\in}\{0,1\}$ with the sigmoid function $x_i = \sigma(y_i) = (1 + e^{-\beta y_i})^{-1}$, where we use $\beta = 10$ for the sharper sigmoid. Thus, from (4), we obtain the following continuous formulation:

$$\boldsymbol{y}^* = \mathrm{argmax}_{\boldsymbol{y}} \textstyle\sum_{i\in V} w_i \sigma(y_i) \prod_{j\in V}(1 - \sigma(y_i))^{\mathbf{B}_{ij}}, \tag{5}$$

where the final solution is obtained from the sigmoid function $\forall i{\in}V$, $x_i^* = (1{+}e^{-\beta y_i^*})^{-1}$. Next, we present our new MWIS algorithm.

3.2. The Algorithm

Our MWIS algorithm iteratively seeks an optimal solution of (5), $\boldsymbol{y}^* \in \mathbb{R}^n$. At each iteration, τ, a current solution is updated using the first-order dynamic:

$$\boldsymbol{y}^{(\tau)} = \boldsymbol{y}^{(\tau-1)} + \Delta\tau\, \dot{\boldsymbol{y}}^{(\tau-1)}. \tag{6}$$

where $\dot{\boldsymbol{y}}^{(\tau-1)} = \frac{d}{d\tau}\boldsymbol{y}^{(\tau-1)}$. For every element of $\boldsymbol{y}$, our goal is to estimate $\dot{y}_i$, such that the final solution, $\boldsymbol{y}^*$, is the maximizer of the objective function of (5). The objective of (5) can be specified as $\sum_i w_i h_i$, where $h_i{=}\sigma(y_i)\prod_{j\in V}(1{-}\sigma(y_i))^{\mathbf{B}_{ij}}$. Thus, it is straightforward to show that the first order dynamic in (6) maximizes $\sum_i w_i h_i$ iff $\sum_i w_i \dot{h}_i \geq 0$, in every iteration τ. From this condition and the definition of h_i, we obtain:

$$\dot{y}_i = \big(1 - \sigma(y_i)\big) w_i h_i - \textstyle\sum_j \mathbf{B}_{ij}\sigma(y_j) w_j h_j, \tag{7}$$

which is used in (6) to obtain the next iterative solution, until convergence. Our algorithm is summarized in Alg. 1.

Algorithm 1: MWIS algorithm

```
Input: graph $G$
Output: MWIS of $G$
Initialize randomly $\boldsymbol{y}^{(0)}$ with $y_i^{(0)} \in \{-1,1\}$;
Compute $\forall i \in V$, $h_i{=}\sigma(y_i)\prod_{j\in V}(1{-}\sigma(y_i))^{\mathbf{B}_{ij}}$;
Compute $\dot{\boldsymbol{y}}^{(0)}$ as in Eq. (7);
while $\|\dot{\boldsymbol{y}}^{(\tau)}\|_2 > 0$ do
    $\Delta\tau \leftarrow LineSearch(\boldsymbol{y})$ ;
    $\boldsymbol{y}^{(\tau+1)} \leftarrow \boldsymbol{y}^{(\tau)} + \Delta\tau\dot{\boldsymbol{y}}^{(\tau)}$ ;
    $\forall i \in V$, $h_i^{(\tau+1)}{=}\sigma(y_i^{(\tau+1)})\prod_{j\in V}(1{-}\sigma(y_i^{(\tau+1)}))^{\mathbf{B}_{ij}}$;
    Update $\dot{\boldsymbol{y}}^{(\tau+1)}$ as in Eq. (7) ;
end
$\boldsymbol{y}^* = \boldsymbol{y}^{(\tau+1)}$;
return $\forall i \in V$, $x_i^* = \sigma(y_i^*)$
```

Theoretical analysis of our algorithm is deferred to Appendix, where we present a proof that Alg. 1 converges to a maximum. From (7), it is easy to show that the complexity of Alg. 1 is $O(n^2)$.

4. Learning Soft Constraints

This section presents our Step 3. As explained in Sec. 3, we conduct multitarget tracking by separately solving the MWIS of each independent subgraph of G, and then link tracklets of the resulting MWIS's into distinct tracks. This procedure can be done online, since every independent subgraph, by construction of G, corresponds only to a pair of consecutive frames. Thus, after solving the MWIS of independent subgraph $G^{(t)}$, we link tracks estimated from previous frames to tracklets of the MWIS of $G^{(t)}$, and thus progressively keep building longer tracks. It is reasonable to expect that the accumulated evidence of statistical appearance, motion, and contextual properties of the targets will help in associating new object detections to the existing tracks. Since data association is controlled by the distance metric, M, and pairwise contextual constraints, $\mathbf{B}$, we seek to learn these parameters from previously tracked instances, as explained in the sequel.

4.1. Distance Metric Learning

From (1), similarity between two object detections (or the weight associated with a tracklet) is defined as a function of the Mahalanobis distance, parameterized by matrix M. We use the well-known large margin nearest neighbor framework to compute M [21]. M is learned to make detections within the same track become closer to each other in the feature space than detections from different tracks. This is formalized as:

$$\begin{aligned} M^* = \arg\min_M \textstyle\sum_{\mathcal{T}_k} \big[\sum_{i,j\in\mathcal{T}_k} (\boldsymbol{z}_i - \boldsymbol{z}_j)^{\mathrm{T}} \mathbf{M} (\boldsymbol{z}_i - \boldsymbol{z}_j) \\ - \textstyle\sum_{i'\in\mathcal{T}_k, j'\notin\mathcal{T}_k} (\boldsymbol{z}_{i'} - \boldsymbol{z}_{j'})^{\mathrm{T}} \mathbf{M} (\boldsymbol{z}_{i'} - \boldsymbol{z}_{j'}) \big], \end{aligned} \tag{8}$$

where the sums are limited to go over k nearest neighbors ($k = 10$). To solve Eq. 8, we use the fast algorithm of [21].

4.2. Pairwise Spatiotemporal Context

We relax the adjacency matrix of G, $\mathbf{B}$, from binary to real values, $\mathbf{B}_{ij} \in [0, 1]$, to account for pairwise spatiotemporal relationships between the tracks. Most importantly, from (4), the relaxation of $\mathbf{B}$ *does not* affect the hard constraints, i.e., the solution of (4) remains MWIS, but introduces additional soft constraints. To this end, we make the assumption that all pairs of objects in the scene have correlated motions. As we demonstrate in our experiments, this additional contextual information improves multitarget tracking. Below, we explain how to relax $\mathbf{B}$.

We consider two cases. Let $i^{(t)}$ and $j^{(t)}$ be a pair of tracklets that are connected by an edge in the graph $G^{(t)}$, $(i^{(t)}, j^{(t)}) \in E^{(t)}$. Then, we keep $\mathbf{B}_{ij} = 1$, as before, to prevent illegal tracks in the MWIS solution. In the second case, $i^{(t)}$ and $j^{(t)}$ are not connected in G, and thus could be included in the MWIS solution. We reason that both $i^{(t)}$ and $j^{(t)}$ should not be members of the MWIS if there is no previous statistical evidence of co-existence of tracks $\mathcal{T}_i^{(t)}$ and $\mathcal{T}_j^{(t)}$ that are constructed by time t, and that end at $i^{(t)}$ and $j^{(t)}$, respectively. Intuitively, if $\mathcal{T}_i^{(t)}$ and $\mathcal{T}_j^{(t)}$ are correlated up to frame t, they are likely to remain correlated from t to $t+1$ if their respective end-tracklets $i^{(t)}$ and $j^{(t)}$ are a good solution. This correlation is estimated, as follows. Let $\boldsymbol{v}_i^{(t)}$ denote the displacement from t to $t+1$ of the moving object corresponding to tracklet $i^{(t)}$, and, similarly, $\boldsymbol{v}_j^{(t)}$ denote the displacement of the moving object corresponding to tracklet $j^{(t)}$. We estimate the 10-bin histogram, H_{ij}, of the $\theta_{ij}^{(t)} = \angle(\boldsymbol{v}_i^{(t)}, \boldsymbol{v}_j^{(t)})$ values during the co-occurrence of $\mathcal{T}_i^{(t)}$ and $\mathcal{T}_j^{(t)}$ in the video, and compute $\mathbf{B}_{ij}$ as

$$\mathbf{B}_{ij} = \begin{cases} 1 & , \text{ if } (i^{(t)}, j^{(t)}) \in E, \\ 1 - H_{ij}(\theta_{ij}^{(t)}) & , \text{ if } (i^{(t)}, j^{(t)}) \notin E. \end{cases} \tag{9}$$

Note that if $\boldsymbol{v}_i^{(t)}$ and $\boldsymbol{v}_j^{(t)}$ do not follow a similar motion pattern, as estimated by time t, then $H_{ij}(\theta_{ij}^{(t)})$ will be close to 0. Then, $\mathbf{B}_{ij}$ will be close to 1, which practically prevents $i^{(t)}$ and $j^{(t)}$ to be in the MWIS solution together. Conversely, if there is a strong statistical evidence across t frames that $\boldsymbol{v}_i^{(t)}$ and $\boldsymbol{v}_j^{(t)}$ are correlated, then $H_{ij}(\theta_{ij}^{(t)})$ will be close to 1. Then, $\mathbf{B}_{ij}$ will be close to 0, which allows both $i^{(t)}$ and $j^{(t)}$ to be in the MWIS solution. In this way, we compute $\mathbf{B}_{ij}$ for all pairs of tracklets of $G^{(t)}$.

5. Handling Long-Term Occlusions

This section presents our Step 4. We extend our method to iteratively find good tracks under long-term occlusions. From the initial set of tracks, obtained by the MWIS algorithm, we first form a new graph, where nodes represent pairs of tracks; weights of nodes represent the average similarity between detections of the two corresponding tracks, given by (1); and edges connect two nodes if the corresponding four tracks share a detection. Then, we find the MWIS of the new graph. The resulting MWIS contains longer mergers of the input smaller tracks. In the next iteration, we again construct a new graph from all the tracks present in the previous MWIS solutions, and find the MWIS of that graph. We also update $\mathbf{M}$ and $\mathbf{B}$ in each iteration, as explained in Sec. 4. The iterations are stopped when the MWIS result does not change.

6. Results

We use five challenging datasets for quantitative evaluation: ETHZ Central [13], TUD Crossing [1], i-Lids AB [10], UBC Hockey [16], and ETHZ Soccer [3]. Videos in these datasets are taken with both static and moving cameras. Targets are seen from varying viewpoints, and under occlusion. Targets also perform different types of movements. In addition, we have compiled our own street-scene dataset of 10 videos, each 2min long, for our qualitative evaluation. Our dataset presents a wide range challenges: cluttered background, occlusion, non static camera, and change of scale. It also complements the above benchmarks, because it provides scenes with objects of different classes, such as bicycles, cars and pedestrians, co-occurring and interacting in the videos. This dataset is available on our website.

We use CLEAR MOT [2, 3] metrics for evaluation. CLEAR MOT consists of: precision—intersection over union of bounding boxes, and accuracy—composed of false negative rate, false positive rate, and number of ID switches.

The steps of our approach are evaluated by starting from a default variant, and then varying one module at a time. The default variant uses the part-based object detector of [7], and LMNN approach of [21] for distance metric learning. Evaluation is conducted on the aforementioned five datasets, and average results are reported in Table 1. Specifically, we run the following four types of experiments.

Dataset	Prec.	Accur.	False Neg.	False Pos.	ID Switch	Run Time
Default	**69.0%**	**81.12%**	**15.5%**	**1.88%**	**1.2**	**44.5 s**
Exp 1.a	67.2%	79.2%	18.23%	1.71%	1.5	41.1 s
Exp 1.b	66.4%	78.54%	19.45%	2.04%	1.5	39.2 s
Exp 2.a	64.1%	76.35%	22.8%	4.21%	3.4	32.8 s
Exp 2.b	67.9%	79.7%	17.71%	2.65%	1.5	40.6 s
Exp 3.a	68.2%	80.3%	18.2%	1.95%	1.2	47.8 s
Exp 3.b	67.9%	79.7%	19.6%	2.15%	1.5	56.2 s
Exp 3.c	66.4%	78.24%	20.45%	2.65%	2.1	44.5 s
Exp 4	54.0%	68.27%	26.4%	6.78%	10.8	34.6 s

Table 1. Average CLEAR MOT [2] results on 5 datasets for evaluating the steps of our approach.

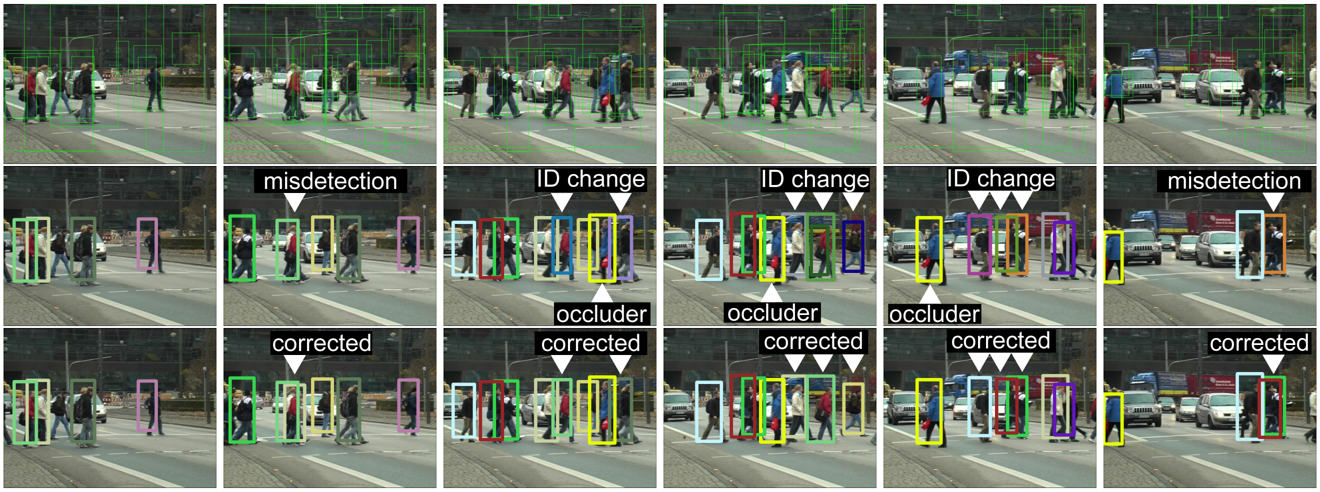
Exp 1: We test the influence of input object detection on performance, by replacing the default part-based detector [7] with the ISM detector [12] (Exp 1.a), and with the HOG detector [5] (Exp 1.b). Table 1 shows the tradeoff between speed and accuracy, where the part-based detector [7] takes longer times, but leads to better tracking performance, on average. **Exp 2:** We evaluate different methods to compute the distance between detected bounding boxes. In addition to the default LMNN approach [21], we also use the simple Euclidean metric where the distance matrix is equal to the identity matrix (Exp 2.a), and also the linear case where the distance matrix is diagonal (Exp 2.b). As can be seen, without distance learning in the case of Exp 2.a, our approach runs faster, but performance significantly decreases, as compared to the default variant. **Exp 3:** Our MWIS algorithm is compared to the maximum weighted clique (MWC) approach of [18] (Exp 3.a), and the iterated tabu search (ITS) of [17] (Exp 3.b). As can be seen, all three methods provide similar average performances. However, our approach is faster. This is because MWC transforms our sparse original graph into highly connected complement graph, which increases complexity. Also, ITS tries to maximize the objective function while eliminating one constraint at a time, whereas we simultaneously consider all constraints, and thus the convergence rate of ITS. For Exp 3.c, we only use the binary version of matrix $\mathbf{B}$. From Table 1, accounting for context improves our tracking results. **Exp 4:** We test the influence of our Step 4, i.e., merging smaller into large tracks to overcome long-term occlusions. Table 1(Exp 4) shows that ID switches decrease dramatically. This demonstrates that most tracks have been merged correctly, and that the system has recovered from occlusions and missed detections.

Dataset	Prec.	Accur.	False Neg.	False Pos.	ID Switch
Central	**72.0%**	**74.2%**	**21.7%**	**0.7%**	**0**
Central[3]	70.0%	72.9%	26.8%	0.3%	0
Central[13]	66.0%	33.8%	51.3%	14.7%	5
Hockey	**60.0%**	**79.7%**	**19.5%**	**1.1%**	**0**
Hockey[3]	57.0%	76.5%	22.3%	1.2%	0
Hockey[16]	51.0%	67.8%	31.3%	0.0%	11
i-Lids	**70.0%**	**78.6%**	**19.4%**	**1.5%**	**1**
i-Lids[3]	66.0%	76.0%	22.0%	2.0%	2
i-Lids[10]	-	68.4%	29.0%	13.7%	-
i-Lids[22]	-	55.3%	37.0%	22.8%	-
Crossing	**73.0%**	**85.9%**	**10.8%**	**1.2%**	**2**
Crossing[3]	71.0%	84.3%	14.1%	1.4%	2
Soccer	**70.0%**	**87.2%**	**6.1%**	**4.9%**	**3**
Soccer[3]	67.0%	85.7%	7.9%	6.2%	4

Table 2. CLEAR MOT [2] results on 5 datasets. Our results are in the top row for each dataset (in bold).

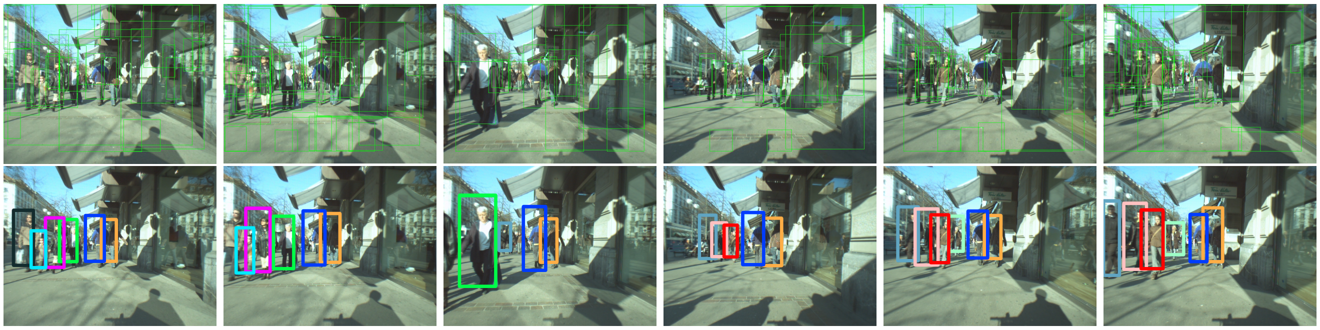
We use the following three competing approaches and datasets for comparison: (i) Coupled detection and trajectory estimation of [13] on ETH Central with provided ground-truth trajectories; (ii) Boosted particle filter of [16] on UBC Hockey; and (iii) Hierarchical data association of [10] on i-Lids. For comparison, we employ the same object detectors as the competing approaches. Specifically, ISM object detector [12] is used for ETH Central, TUD crossing, and UBC Hockey, and the HOG detector [5] is applied to i-Lids and Soccer. The detectors are implemented in their generic, publicly available, pre-trained versions, i.e., they are not specifically trained for any test sequence, unlike [16]. We use only 2D visual cues, and do not assume any prior knowledge about the video contents, such as, e.g., ground plane, camera calibration, or entry/exit zones, used in [13, 10]. The comparison results are reported in Table 6. As can be seen, our multitarget tracking has high precision and accuracy. Errors occur when a target person is: (i) very close to other targets in the ETH Central, TUD Crossing sequences; (ii) sitting in the ETH Central videos; or (iii) partially out of the field of view in the i-Lids videos. ID switches in i-Lids happen mainly when a target person is occluded for a long time (e.g., by a pillar), and a new track is initialized for the person's reappearance. For sports sequences, ID switches are more often, because players in the videos have very similar appearance and motion properties. From Table 6, we outperform the competing approaches on all datasets.

For qualitative evaluation, we use three datasets: TUD Crossing, ETH moving vehicle [6], and our own dataset. Also, in all qualitative evaluations described below, we apply the part-based object detector of [7]. Fig. 3 shows our results at different steps of our approach, on a sequence from TUD Crossing. The top row shows object detection responses. The middle row shows our tracking results before Step 4. The bottom row presents the tracking results after our Step 4. As can be seen, after Steps 1-3, many tracks are cut short, due to missed detections, or occlusion from the blue person crossing in the opposite direction. The bottom row shows that we recover from these errors after Step 4. Next, Fig. 4 shows in the top row a sequence from the ETH moving vehicle dataset and object detection responses, and in the bottom row our final results. As can be seen, our approach performs well under camera motion, and addresses well relatively large scale changes of pedestrians. Finally, Fig. 5 shows our tracking results on a sequence from our street-scene dataset. We apply the part-based detector of [7] to detect pedestrians, bicycles and cars



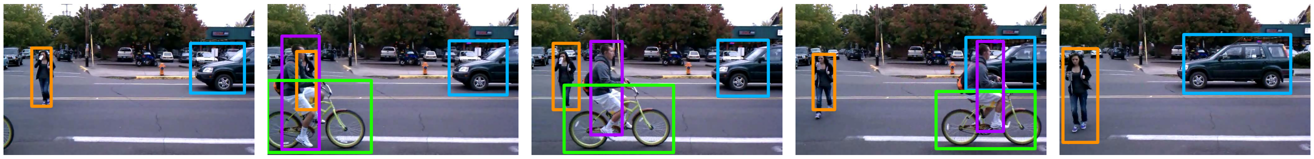
Frame 1 Frame 24 Frame 50 Frame 70 Frame 98 Frame 122

Figure 3. Qualitative results on a TUD Crossing sequence that contains the occlusion from the blue pedestrian (occluder detected in the red box) crossing in the opposite direction from the crowd. The top row shows responses of the part-based object detector of [7]. The middle row shows our tracking results before Step 4, and the bottom row, after Step 4. We see that Step 4 corrects tracking errors due to the long-term occlusion.



Frame 224 Frame 230 Frame 243 Frame 263 Frame 278 Frame 286

Figure 4. Qualitative results on a sequence from the ETH moving vehicle dataset. The top row shows object detection responses of [7], and the bottom row shows that we can handle occlusion, moving camera and change of scale.



Frame 1165 Frame 1173 Frame 1177 Frame 1187 Frame 1210

Figure 5. Qualitative results on a sequence of our dataset. Car, bike and pedestrian detections are put in the same bag of detections. We are able to track different objects simultaneously under occlusion.

in the same video. All detection responses of cars, bikes, and pedestrians are put in the same bag of detections. Fig. 5 shows that despite occlusion, our system is still able to track different objects simultaneously. Our qualitative evaluation on 10 videos demonstrates that capturing spatiotemporal interactions between objects of different classes helps tracking each object.

7. Conclusion

We have presented a tracking-by-detection approach, where associating object detections with tracks is formulated as finding the MWIS of a graph of tracklets. A new MWIS algorithm, and its theoretical analysis have been presented. The MWIS formulation is capable of explicitly encoding both soft and hard spatiotemporal interactions between objects in a unified manner. Our main contributions include: generalizing bipartite one-to-one matching, used in prior work for multitarget tracking, to a more powerful framework, that of MWIS; and accounting for long-term motion correlations among the tracks. We outperform competing approaches on challenging benchmarks, in terms of the CLEAR MOT metrics.

Appendix

This section presents a theoretical analysis of Alg. 1.

Theorem 1 *The objective function of Eq. (5) does not decrease under the dynamic defined by Eq. (7).*

Proof: We prove that when $\dot{y}_i$ is computed as in Eq. (7), we have $\sum_i w_i \dot{h}_i \geq 0$. Define $x_i = \sigma(y_i)$. From the definition of h_i (see Sec. 3.2), we have $\dot{h}_i = \beta h_i \Big((1 - x_i)\dot{y}_i - \sum_j \mathbf{B}_{ij} \dot{y}_j x_j \Big)$. It follows $\sum_i w_i \dot{h}_i {=} \sum_i w_i \beta h_i \Big((1{-}x_i)\dot{y}_i {-} \sum_j \mathbf{B}_{ij} \dot{y}_j x_j \Big) {=} \boldsymbol{u}^{\mathrm{T}} \mathbf{A} \dot{\boldsymbol{y}}$, where $u_i = w_i h_i$, and the auxiliary matrix $\mathbf{A}$ has the following elements: $\mathbf{A}_{ij}{=}1{-}x_i$, if $i{=}j$, else, $\mathbf{A}_{ij}{=}{-}x_j$, if $(i,j){\in}E$, and $\mathbf{A}_{ij}{=}0$, otherwise. Thus, by computing $\dot{\boldsymbol{y}}{=}\mathbf{A}^{\mathrm{T}}\boldsymbol{u}$, as in Eq. (7), we obtain $\sum_i w_i \dot{h}_i = \boldsymbol{u}^{\mathrm{T}} \mathbf{A} \mathbf{A}^{\mathrm{T}} \boldsymbol{u} \geq 0$. □

Corollary 1 *Strict inequality $\sum_i w_i \dot{h}_i > 0$ cannot be achieved, since $\mathbf{A}\mathbf{A}^T$ is not positive definite.*

Proof: We prove that $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is not positive definite. The MWIS contains at least one node, e.g., $x_i = 1$. It follows, $\forall j \in V$, $(i,j) \in E$, $x_j = 0$. Then, all the elements of ith row of $\mathbf{A}$ are zero, i.e., $\mathbf{A}$ does not have the full rank. Consequently, at least one of the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is zero. □

Theorem 2 *Alg. 1 converges to a local maximum.*

Proof: Since $\forall i$, $x_i = \sigma(y_i) : \mathbb{R} \rightarrow [0,1]$, it follows $\forall i$, $h_i : \mathbb{R} \rightarrow [0,1]$. Consequently, $\sum_i w_i \dot{h}_i \leq \boldsymbol{w}^{\mathrm{T}} \mathbf{1}$ where $\mathbf{1}$ is the vector of 1's. Since $\sum_i w_i \dot{h}_i$ always increases (see Th.1) and $\sum_i w_i \dot{h}_i$ is upper bounded, Alg. 1 converges. The algorithm stops when the gradient $\left\| \dot{\boldsymbol{y}}^{(t)} \right\|_2 = 0$, i.e., in a local maximum. □

Acknowledgment

The support of the National Science Foundation under grant NSF IIS 1018490 is gratefully acknowledged.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

[2] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *J. Image Video Process.*, 2008:1–10, 2008.

[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[4] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, 2010.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[6] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.

[7] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models. In *CVPR*, 2009.

[8] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, pages 260–267, 2006.

[9] H. Grabner, J. Matas, L. Van Gool, and P. Cattin. Tracking the invisible: Learning where the object might be. In *CVPR*, 2010.

[10] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[11] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010.

[12] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vision*, 77(1-3):259–289, 2008.

[13] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

[14] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, pages 2953–2960, 2009.

[15] Y. Li and R. Nevatia. Key object driven multi-category object recognition, localization and tracking using spatio-temporal context. In *ECCV*, 2008.

[16] K. Okuma, A. Taleghani, N. D. Freitas, O. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

[17] G. Palubeckis. Iterated tabu search for the unconstrained binary quadratic optimization problem. *Informatica*, 17(2):279–296, 2006.

[18] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *PAMI*, 29(1):167–172, 2007.

[19] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *CVPR*, 2006.

[20] K. Shafique, M. W. Lee, and N. Haering. A rank constrained continuous formulation of multi-frame multi-target tracking problem. In *CVPR*, 2008.

[21] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *ICML*, 2008.

[22] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *Int. J. Comput. Vision*, 75(2):247–266, 2007.

[23] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 2009.

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