Vehicle Tracking in Wide Area Motion Imagery via Stochastic Progressive Association Across Multiple Frames (SPAAM)

Ahmed Elliethy, *Student Member, IEEE* and Gaurav Sharma, *Fellow, IEEE*

Abstract—Vehicle tracking in Wide Area Motion Imagery (WAMI) relies on associating vehicle detections across multiple WAMI frames to form tracks corresponding to individual vehicles. The temporal window length, i.e., the number M of sequential frames, over which associations are collectively estimated poses a trade-off between accuracy and computational complexity. A larger M improves performance because the increased temporal context enables the use of motion models and allows occlusions and spurious detections to be handled better. The number of total hypotheses tracks, on the other hand, grows exponentially with increasing M, making larger values of M computationally challenging to tackle. In this paper, we introduce SPAAM an iterative approach that progressively grows M with each iteration to improve estimated tracks by exploiting the enlarged temporal context while keeping computation manageable through two novel approaches for pruning association hypotheses. First, guided by a road network, accurately co-registered to the WAMI frames, we disregard unlikely associations that do not agree with the road network. Second, as M is progressively enlarged at each iteration, the related increase in association hypotheses is limited by revisiting only the subset of association possibilities rendered open by stochastically determined dis-associations for the previous iteration. The stochastic dis-association at each iteration maintains each estimated association according to an estimated probability for confidence, obtained via a probabilistic model. Associations at each iteration are then estimated globally over the M frames by (approximately) solving a binary integer programming problem for selecting a set of compatible tracks. Vehicle tracking results obtained over test WAMI datasets indicate that our proposed approach provides significant performance improvements over state of the art alternatives.

Index Terms—Wide area motion imagery, vehicle tracking, association hypotheses pruning, binary integer programming, vector road network

I. INTRODUCTION

AERIAL photography has come a long way since 1858, when the first documented aerial photograph was captured from onboard a balloon in France by Gaspard-Félix Tournachon [2]. Today satellite, aircraft, drone, and balloon based aerial still-images are commonplace. Additionally, newer platforms offer motion imagery with rich spatio-temporal information that enables a host of new applications. We focus particularly on urban area wide area motion imagery (WAMI) that offers high resolution image sequences covering large field of view (city-scale) within each frame, at temporal rates of 1-2 frames per second (fps) [3]–[5]. In this setting, we consider the problem of tracking the many vehicles present in the field of view.

Tracking is an extensively researched problem (see [6] for a recent survey). Our discussion focuses specifically on methods applicable in the WAMI setting, which poses a number of unique challenges. Typically vehicles span only a few pixels and there is little spatial detail discriminating individual vehicles from each other based on appearance. Therefore, vehicles are usually represented only by their spatial locations. Moreover, the potentially large number of vehicles included within the spatial coverage of typical WAMI frames make tracking computationally demanding. The difficulty of the problem is compounded by the relatively low 1-2 fps (in contrast with 30 fps full motion video) temporal resolution in WAMI, due to which the spatial extent of a vehicle in adjacent frames does not typically overlap and can, in fact, be a fair distance apart. Further complexity and challenges are introduced by spurious and missed detections caused by imperfections in the models used for detecting vehicles against the background and due to occlusions of vehicles by overbridges, trees, buildings, etc.

The aforementioned challenges mandate specific approaches for vehicle tracking in WAMI. Specifically, tracking is performed by first detecting the *spatial locations* of vehicles (for instance, based on their movement relative to the background) and then forming vehicle tracks by associating detections that are presumed to correspond to the same vehicle over the set of WAMI frames. To obtain the associations, hypothesis tracks are first assigned costs that penalize deviation from a motion model that typifies the common behavior of moving vehicles. Detections are then associated into tracks with the objective of minimizing the total cost for the tracks, under the constraint that a vehicle detection is assigned to at most one track. The size of the temporal context, i.e. the number of (adjoining) frames, used in formulating the motion model poses an inherent trade-off. Costs that assess the plausibility of the vehicle motion postulated by a hypothesis track are better assessed using motion models over a larger temporal context. However, the number of possible tracks increases

A. Elliethy was with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (e-mail: ahmed.s.elliethy@rochester.edu).

G. Sharma is with the Department of Electrical and Computer Engineering, Department of Computer Science, and Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14627, USA (e-mail: gaurav.sharma@rochester.edu).

exponentially with the number of frames, which makes it challenging to estimate the set of tracks with minimum total cost for a larger temporal context.

Existing tracking techniques handle the complexity introduced by the larger temporal context by one of two common approaches. The first approach solves the minimum cost tracking problem exactly but with a simplified track cost that is decomposable into a sum of independent costs for pairwise associations between detections in adjacent frames that make up the track. A network is formed by arranging the detections as nodes and feasible associations between detections in adjacent frames as edges with corresponding costs. Network paths consisting of a sequence of connected edges then correspond to a feasible track whose cost is the sum of the costs of the edges in the path. The tracking problem is then posed as a minimum cost maximum flow estimation problem, which is solved exactly using one of several network flow optimization algorithms [7]–[11]. Online tracking algorithms [12], [13] that build the tracks progressively are also frequently based on the simplified cost approach. A bipartite graph is constructed for each pair of successive frames, and pairwise associations across the frames are estimated by finding the minimum cost bipartite matching. Simplified pairwise independent costs are also the basis of hierarchical track estimation methods [14]–[17]. Hierarchical approaches first estimate short tracklets, then hierarchically link estimated tracklets together to form longer tracks. In [14]–[16], the Hungarian algorithm [18] is used to estimate pairwise associations to estimate tracklets with minimum cost. In [17], a network flow based algorithm is employed for both estimating tracklets and to link them using novel learnt tracklet discriminative metrics. In the WAMI setting, vehicle detections are only represented by their spatial locations, therefore it is beneficial to use costs based on motion models, such as constant velocity/acceleration, which rely on larger spatial context and do not meaningfully map to (independent) pairwise association costs that are required for the aforementioned algorithms.

The second class of approaches tackles the complexity introduced by the larger temporal context by utilizing efficient methods to solve the minimum cost tracking problem approximately, without limiting the focus to independent pairwise decomposable costs. A number of iterative and/or hierarchical techniques are representative of the recent developments in this category. In [19], [20], an approximate solution to the tracking problem is obtained by iteratively re-estimating pairwise associations between two successive frames while keeping other associations fixed. In each iteration, effective "costs" for the pairwise associations being re-estimated are obtained by collapsing the costs over the associations that are kept fixed and the pairwise associations are re-estimated either as binary variables via the Hungarian algorithm [19] or as real-valued "soft" assignments computed via a rank-one tensor factorization [20] which are finalized upon completion of the iterations. Motion models over larger temporal context can be utilized in the aforementioned approximate methods. The computational tractability obtained through the approximate approaches is, however, at a cost; the methods only ensure convergence to locally optimal solutions to the minimum

cost tracking problem. Furthermore, because each iteration updates only the associations between adjacent frames often the methods converge to a weak local optima and the benefit of the cost over the larger temporal context is not fully realized. Alternative approaches, specifically focusing on a constant velocity motion model in a three frame context are proposed in [21], [22]. In [21], tracklets are first estimated over a sliding three frame context using an approximate combinatorial optimization technique, consistent tracklets between adjacent three frame windows are merged and an enlarged four frame context is used to resolve conflicts; finally a min-cost network flow based global optimization is used to link together tracklets with the objective of overcoming fragmentation due to long-term occlusions. This bottom-up procedure suffers from the limitation that errors introduced in the tracklet estimation stage cannot be corrected in subsequent steps. To better maintain the global temporal context, the approach in [22] forms an auxiliary network whose nodes represent ordered pairs of possible associations of detections from adjacent frames. The nodes are joined together by links with costs determined by the constant velocity model for each link's 3-frame context. Because the same detection is replicated multiple times in the nodes formed by the ordered pairs, the minimum cost tracking formulation on the auxiliary network requires consistency constraints to ensure each detection is assigned only to one track. While the problem formulation does not directly correspond to the well-studied network flow problem due to these constraints, an approximate solution can be obtained by iteratively using network flow optimization techniques in combination with Lagrangian relaxation [23]. The constraints are mapped to costs included in the objective function using Lagrange multipliers that are progressively adapted through the iterations to provide a final solution consistent with the constraints. The work in [24] considers an alternative auxiliary network based formulation where combinatorial expansion of the local context allows alternative constant velocity costs to be computed between nodes separated by a number of intervening frames. Using a shortest path algorithm iteratively in a greedy fashion, tracks are then estimated one at time followed by elimination of nodes from the network that have been consumed by estimated tracks (which ensures that nodes are used only in at most one estimated track). The advantage of global context in these methods comes at the cost of considerable increase in complexity because of the increased size of the auxiliary network compared with the network used in the traditional network flow formulation of tracking with decomposable costs. In particular, for the large scale of WAMI tracking problems, these approaches are not directly applicable[1].

In this paper, we propose SPAAM an alternative computationally-efficient, hierarchical, and iterative approach for solving the WAMI tracking problem over a relatively large temporal context without requiring track costs that are decomposable as a sum of independent pairwise association costs.

[1]To solve a tracking problem over N frames with D detections per frame, the technique in [22] generates an auxiliary network with $(N-1)D^2$ nodes and $(N-2)D^3$ edges compared with $N \times D$ nodes and $(N-1)D^2$ edges in the traditional flow networks.

In each SPAAM iteration, the entire set of available WAMI frames are partitioned into smaller temporal windows of contiguous frames over which associations are first estimated; associations are then hierarchically linked across the windows to obtain the overall associations over the entire set of WAMI frames. The temporal windows are expanded in successive iterations to benefit from a larger temporal context. To handle the relatively large temporal windows, two innovations are proposed to limit the number of association hypotheses and to render the approach computationally tractable for the relatively large number of tracks observed in WAMI. First, guided by a pixel accurate co-registered road network (RN) with the WAMI frames, we disregard unlikely associations that do not agree with the RN; only association hypotheses that link *road reachable* locations in adjacent frames are considered. Second, as the temporal window size is increased with progressing iterations, the associated increase in association hypotheses is limited by revisiting only a subset of associations. This is accomplished by: (a) stochastically preserving/dissolving associations established at the previous iteration such that the probability of preservation matches the estimated confidence probability that the association is correct, and (b) for the current iteration, maintaining the preserved associations and estimating only the associations remaining open. The confidence of every pairwise association is estimated via a probabilistic model and the estimation of open associations at each iteration is formulated as a binary integer programming (BIP) problem. In comparison with existing techniques, the proposed SPAAM approach provides two key advantages. First, because the proposed approach is not constrained to using costs that are decomposable as a sum of independent pairwise association costs, it can use more meaningful costs for the WAMI setting based on models for vehicular motion. Second, because SPAAM dissolves the (predominantly) low confidence associations and estimates remaining open associations collectively over a larger temporal context, instead of simply re-estimating the associations between an adjacent pair of frames, the method invariably overcomes the local minima problem that plagues prior methods that allowed more general costs. Results obtained with SPAAM demonstrate a significant improvement over the alternative methods benchmarked.

This paper is organized as follows. Section II presents the formulation for the problem of estimating vehicle detection associations over multiple WAMI frames. Section III explains our proposed stochastic progressive multi-frame data association approach. Results and a comparison against alternative methods are presented in Section IV, followed by a discussion in Section V. We conclude the paper in Section VI.

II. Multi-frame association problem formulation

Consider the problem of tracking individual vehicles over a set $\mathcal{I} = \{I_i\}_{i=1}^{N}$ of N WAMI frames, where the WAMI frames in $\mathcal{I}$ are co-registered with a vector road network defined in a common reference[2] coordinate system $\mathcal{R}$ (for example

[2]Specifically, we use a 2D Cartesian map coordinate system obtained via azimuthal orthographic map projection (AOMP) [25] of the spherical latitude and longitude coordinates.

N	Total number of WAMI frames
M	Size of temporal window
$\mathcal{I}$	Sequence of N WAMI frames $I_1, I_2, \ldots, I_N$.
$\mathcal{R}$	A common coordinate system for all frames in $\mathcal{I}$
D_i	Number of vehicle detections (VDs) in the ith WAMI frame
z_j^i	The location of the jth VD detected in the ith WAMI frame represented in the coordinate system $\mathcal{R}$
G	Trellis graph
v_j^i	Node in the trellis graph that G represents the VD z_j^i
$\mathcal{V}_i$	Set of nodes in the ith frame
$V(G)$	Set of nodes of G
$\varepsilon_{a,b}^i$	Trellis graph (directed) edge that links node v_a^i to v_b^{i+1}
$E(G)$	Set of edges of G
$s(G)$	First frame index represented in G
$e(G)$	Last frame index represented in G
$\mathcal{P}(G)$	Set of paths that can be represented on G
$\mathcal{G}$	Edge induced sub-graph
$\vartheta(a,b)$	Minimum travel distance from a to b on the road network
P	Track
$\mathcal{A}$	Adjacency set
$\mathcal{C}\left(v_j^i\right)$	Set of all nodes reachable from v_j^i on the road network
t	Iteration index
ℓ	Hierarchy level index
w	Temporal window index
k	Track index
$(.)^{(t)}$	Value of a variable in the tth iteration
$(.)^{(t,\ell)}$	Value of a variable at the ℓth hierarchy level in the tth iteration

Key notation/symbols used in the paper

using [26]–[28]). We adopt a tracking by detection paradigm: vehicle locations are detected in the individual frames and the goal is to associate the detections of a vehicle over the entire set of WAMI frames to form a *track* per vehicle. Frame I_i contains D_i vehicle detections (VDs) and $z_j^i = (x_j^i, y_j^i)$ denotes the spatial location of the jth VD in the ith frame in the coordinate system $\mathcal{R}$.

To describe the problem and the solution approach, we use a 2D trellis graph G constructed as follows for a temporal window of WAMI frames, whose starting and ending frame numbers we denote by $s(G)$ and $e(G)$, respectively. Along the horizontal axis, nodes in the trellis graph G lie along $T(G) = (e(G) - s(G)) + 1$ discrete positions corresponding to the frame instances from $s(G)$ through $e(G)$ arranged in increasing order from left to right. At the horizontal position for frame i, along the vertical axis, $D_i + 1$ distinct nodes $\mathcal{V}_i = \left\{ v_0^i, v_1^i, \ldots, v_{D_i}^i \right\}$ are placed. For $j = 1, 2, \ldots D_i$, the node v_j^i represents the jth VD in the ith frame (at spatial location z_j^i in $\mathcal{R}$) and v_0^i represents a dummy node that is introduced to account for missed-detections and for identifying spurious detections (as described later). Edges are introduced in the trellis graph linking nodes across adjacent frames[3] that can correspond to detections of the same vehicle, or that link a VD to a dummy node. We denote by $\varepsilon_{a,b}^i$ the (directed) edge that links v_a^i with v_b^{i+1} and by $\mathcal{A}\left(v_a^i\right)$ the adjacency set for v_a^i that contains all VD nodes in the $(i+1)^{\text{th}}$ frame linked with v_a^i. In the absence of additional information, $\mathcal{A}\left(v_a^i\right)$ contains all other nodes in the $(i+1)^{\text{th}}$ frame. In practice, however, physical considerations rule out a number of possibilities and the adjacency set $\mathcal{A}\left(v_a^i\right)$ is much smaller than the full set of nodes in the $(i+1)^{\text{th}}$ frame. For example, no edges are permissible between VD nodes whose spatial locations are

[3]The trellis graph is characterized by the property that the only permitted edges are those that link nodes across adjacent temporal locations.
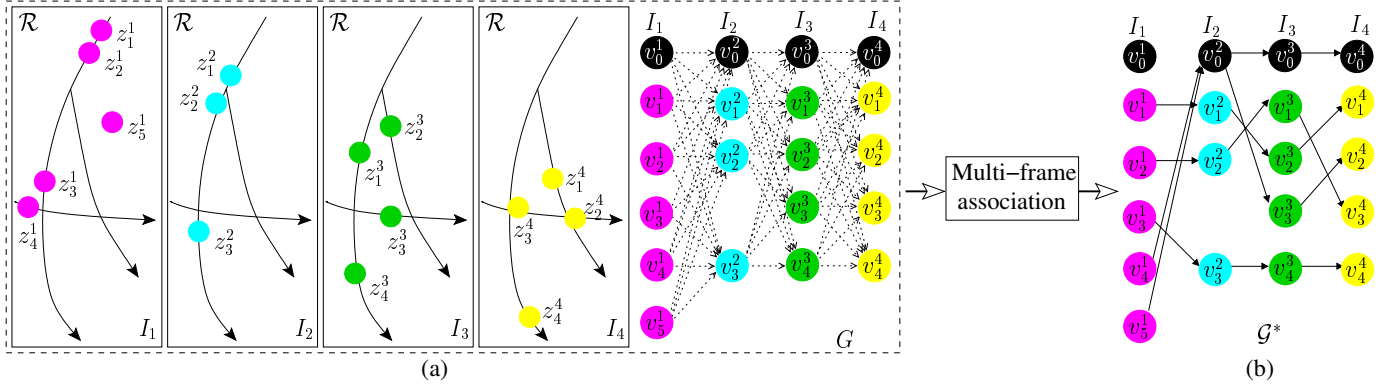
Fig. 1: A schematic illustration of the problem formulation via an example. The input comprises an $N = 4$ frame WAMI sequence $\mathcal{I} = \{I_1, I_2, I_3, I_4\}$ that is co-registered in a common reference coordinate system $\mathcal{R}$. Frame I_i has D_i VDs, with the location of the jth VD denoted by z_j^i. A trellis graph G is obtained by associating discrete horizontal locations with the frames (in sequence) and arranging nodes vertically at the ith horizontal position to represent the VDs in the corresponding frame. Node v_j^i represents the VD at location z_j^i, for $j = 1, 2, \ldots D_i$ and a dummy node v_0^i is introduced to handle missed/spurious detections. The VDs in adjacent frames that can correspond to the same vehicle are connected by directed edges as shown in (a). A hypothesis vehicle track is represented on G by a path that consists of consecutive sequence of nodes from frame $s(G) = 1$ to frame $e(G) = 4$. The goal in tracking is to estimate the best consistent subset of paths on $\mathcal{P}(G)$ out of the set of all possible paths, or to equivalently estimate the best edge induced sub-graph $\mathcal{G}^*$ from G such that any non-dummy node in the sub-graph must belong to exactly one path (track) as shown in (b). Note that, dummy nodes are used for missed-detections as shown in the path $(v_4^1, v_0^2, v_3^3, v_2^4) \in \mathcal{P}(\mathcal{G}^*)$, and for identifying spurious detections as shown in $(v_5^1, v_0^2, v_0^3, v_0^4) \in \mathcal{P}(\mathcal{G}^*)$.

further apart than the maximal distance that a vehicle can traverse in the inter-frame interval. The sets $V(G) = \bigcup_{i=s(G)}^{e(G)} \mathcal{V}_i$ and $E(G) = \left\{\varepsilon_{a,b}^i \,\middle|\, v_b^{i+1} \in \mathcal{A}(v_a^i), v_a^i \in \mathcal{V}_i, v_b^{i+1} \in \mathcal{V}_{i+1}\right\}$ represent the complete set of nodes and edges, respectively, in the graph G. An example trellis graph for a temporal window starting at frame $s(G) = 1$ and ending at frame $e(G) = 4$ is shown in Fig. 1 (a).

Within the temporal support of G, a vehicle track is represented as a sequence $P_k = \left(v_{\zeta_{s(G)}^k}^{s(G)}, v_{\zeta_{s(G)+1}^k}^{s(G)+1}, \ldots, v_{\zeta_{e(G)}^k}^{e(G)}\right)$ of nodes $\zeta_i^k \in \{0, 1, \ldots, D_i\}$ corresponding to the locations of a (single) vehicle in the frames. Specifically, $\zeta_i^k \in P_k$ implies that the vehicle track passes through the location z_j^i if $j \neq 0$ and that the vehicle is not detected in the ith frame if $j = 0$. Alternatively, to simplify notation, we also denote the track as $P_k = \left(v_{k_{s(G)}}^{s(G)}, v_{k_{s(G)+1}}^{s(G)+1}, \ldots, v_{k_{e(G)}}^{e(G)}\right)$ at the cost of some precision.

Because only vehicle locations are available in our setting, for estimating tracks, we proceed as follows. We denote by $\mathcal{P}(G)$ the set of all feasible tracks on G and for each feasible track $P_k \in \mathcal{P}(G)$, we assign a cost $C(P_k)$ that assesses the plausibility of the vehicle movement predicated by the track. The cost decreases monotonically as the corresponding vehicle movement becomes more plausible. The objective of our tracking problem is then formulated as the task of estimating the best, i.e. minimum total cost, consistent subset of tracks from $\mathcal{P}(G)$, where "consistent" means that every non-dummy node belongs to exactly one track. In the graph formulation, the problem of estimating the best consistent subset of tracks is equivalent to the problem of selecting the minimum-cost feasible edge-induced sub-graph $\mathcal{G}^*$ from the set $\mathcal{F}(G)$ of all feasible edge-induced sub-graphs of G. A feasible edge-

induced sub-graph $\mathcal{G} \in \mathcal{F}(G)$ is a graph that has the same set of nodes of G, i.e. $V(\mathcal{G}) = V(G)$, but has a subset of edges from G such that any non-dummy node in $V(\mathcal{G})$ belongs to exactly one track in $\mathcal{G}$ as illustrated in the example shown in Fig. 1 (b). Formally, a feasible edge-induced sub-graph $\mathcal{G} \in \mathcal{F}(G)$ is defined by the constraints

$$\sum_{P_k \in \mathcal{P}(G)} \chi_{\mathcal{P}(\mathcal{G})}(P_k) \, \chi_{P_k}(v_j^i) = 1, \ \forall v_j^i \in V(\mathcal{G}), j \neq 0,$$
$$V(\mathcal{G}) = V(G), \tag{1}$$

where, for any set A, $\chi_A(\cdot)$ is the indicator function, defined as

$$\chi_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases} \tag{2}$$

Our tracking problem is then formulated as

$$\mathcal{G}^* = \arg \min_{\mathcal{G} \in \mathcal{F}(G)} \psi(\mathcal{G}), \tag{3}$$

where

$$\psi(\mathcal{G}) = \sum_{P_k \in \mathcal{P}(\mathcal{G})} C(P_k). \tag{4}$$

We model the global cost C of a track P_k as exponentiated weighted sum of metrics that penalize (a) motion irregularities, (b) road network dis-agreement, and (c) number of dummy nodes along the track. The metrics comprise (non-independent) pairwise association costs estimated over the local context associated with every pairwise association $\varepsilon_{k_i, k_{i+1}}^i$ in the track P_k. Specifically, the cost of the track P_k is defined as in (5), where:

- σ_m, σ_d, σ_θ, and σ_g are nonnegative weighting factors,

$$C\left(P_k\right) = -\exp\left(-\sum_{i=s(G)}^{e(G)} \frac{\sigma_m \Gamma\left(\varepsilon_{k_i,k_{i+1}}^i\right) + \sigma_d R_d\left(\varepsilon_{k_i,k_{i+1}}^i\right) + \sigma_\theta R_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)}{L_p\left(P_k\right)} + \frac{\sigma_g L_e\left(\varepsilon_{k_i,k_{i+1}}^i\right)}{T(G)}\right),\tag{5}$$

- $L_p\left(P_k\right)$ is the length of P_k excluding dummy nodes, i.e., $L_p\left(P_k\right) = |\{v_{k_i}^i | v_{k_i}^i \in P_k,\ k_i \neq 0\}|$,
- $\Gamma\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ is a motion irregularity penalty defined as

$$\Gamma\left(\varepsilon_{k_i,k_{i+1}}^i\right) = 1 - \frac{\Gamma_m\left(\varepsilon_{k_i,k_{i+1}}^i\right) + \Gamma_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)}{2},\tag{6}$$

where $\Gamma_m\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ and $\Gamma_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ are normalized measures (in the rage $[0,1]$) that quantify the similarity, in magnitude and direction, respectively, of the velocity for the ith frame with the velocities for other frames in a temporal window of $(2W+1)$ frames centered about the ith frame. The computation of these terms is illustrated using Fig. 2. The velocity at the jth frame is computed as $\mathbf{v}_{k_j}^j = \left(z_{k_{j+1}}^{j+1} - z_{k_j}^j\right)/\Delta t$ where Δt is the interframe duration, and inspired by [20], the magnitude and direction similarity terms are obtained as

$$\Gamma_m\left(\varepsilon_{k_i,k_{i+1}}^i\right) = \frac{1}{2W} \sum_{\substack{m=-W,\\ W \neq 0}}^{W} \frac{2 \left\|\mathbf{v}_{k_i}^i\right\| \left\|\mathbf{v}_{k_{i-m}}^{i-m}\right\|}{\left\|\mathbf{v}_{k_i}^i\right\|^2 + \left\|\mathbf{v}_{k_{i-m}}^{i-m}\right\|^2},\tag{7}$$

$$\Gamma_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right) = \frac{1}{2W} \sum_{\substack{m=-W,\\ W \neq 0}}^{W} \frac{\left\langle \mathbf{v}_{k_i}^i, \mathbf{v}_{k_{i-m}}^{i-m}\right\rangle}{\left\|\mathbf{v}_{k_i}^i\right\| \left\|\mathbf{v}_{k_{i-m}}^{i-m}\right\|}.\tag{8}$$

The term $\Gamma_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ represents the average of cosines of the angles between the direction of the velocity vector $\mathbf{v}_{k_i}^i$ and the directions of the other velocity vectors in the temporal window. If the motion of the track is smooth within the temporal window, both $\Gamma_m\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ and $\Gamma_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ are close to one and $\Gamma\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ is close to zero. For tracks with dummy nodes, velocities are computed between actual VD locations that become adjacent once the dummy nodes are skipped.

- $R_d\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ and $R_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ are costs penalizing, respectively, the distance and direction disagreement with the road network, computed using the factors illustrated in Fig. 2. Specifically, $R_d\left(\varepsilon_{k_i,k_{i+1}}^i\right) = d_{k_i}^i / r_{k_i}^i$ is the distance $d_{k_i}^i$ of the VD location $z_{k_i}^i$ to the nearest point on the center line of the road, normalized through division by the factor $r_{k_i}^i$ that represents the number of lanes on that road, and $R_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right) = 1 - \cos\left(\theta_{k_i}^i\right)$, where $\theta_{k_i}^i$ is the angle between the velocity vector $\mathbf{v}_{k_j}^j$ and the road (in the direction of travel) at the nearest point on the center line of the road. The terms $R_d\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ and $R_\theta\left(\varepsilon_{k_i,k_{i+1}}^i\right)$ are set to zero when $v_{k_i}^i$ represents a dummy node.

- $L_e\left(\varepsilon_{k_i,k_{i+1}}^i\right) = \chi_{\{0\}}(k_{i+1})$ indicates whether $v_{k_{i+1}}^{i+1}$ is a dummy node or not.

The exponentiation in (5) moderates the cost for each component and the overall cost for a hypothesis track is in the interval $[-1,0)$. Also, the form of (5) ensures that tracks rated poorly for any one of the penalty terms also have an overall low cost.

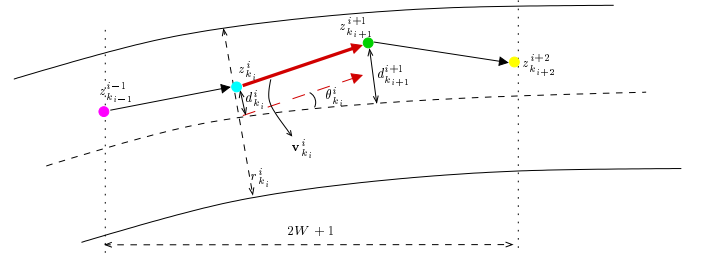

Fig. 2: Schematic illustrating the metrics used for modeling the cost of a track P_k.

III. STOCHASTIC PROGRESSIVE ASSOCIATION ACROSS MULTI-FRAMES (SPAAM)

To track the vehicles in the WAMI image sequence $\mathcal{I}$, we wish to solve (3). For our proposed cost and typical number of frames, this multi-frame association problem is computationally intractable. In particular, the constrained minimization problem in (3) is an instance of BIP, which is NP-*complete* in its general form [29]. Although, a multi-clique formulation of the data association has been proposed and solved in the BIP framework for relatively small numbers of tracks [30], [31], these methodologies do not scale to the large number of vehicles seen in WAMI data. SPAAM, the approach proposed in this paper, solves (3) approximately in a computationally tractable manner.

Operating iteratively and hierarchically, SPAAM estimates associations between adjacent vehicle detection nodes in the trellis graph G to form the estimated vehicle tracks. The tth iteration first estimates associations over non-overlapping windows of $M^{(t)}$ frames each and then hierarchically computes associations across windows. The window size $M^{(t)}$ is progressively enlarged with successive iterations. To keep the complexity from rapidly growing with the increase in window size in successive iterations, only a subset of stochastically determined associations are revisited at each iteration after the first iteration. Specifically, estimated tracks at each iteration are assessed to evaluate a confidence probability for each postulated pairwise association between vehicle detection nodes and, prior to the next iteration, pairwise associations are stochastically retained with a probability matching the estimated confidence probability and disassociated otherwise. The next iteration then only considers prospective associations for the vehicle detection nodes that were stochastically

disassociated while using the larger temporal context and maintaining the associations that were stochastically retained. By this process, SPAAM limits less-useful re-evaluations of association decisions that already have a high confidence and focuses available computation primarily on determining remaining associations with the benefit of a larger temporal context. A road network, accurately co-registered with the WAMI frames in $\mathcal{I}$, provides useful spatial context for SPAAM allowing prospective associations to further be limited to road-reachable pairs of vehicle detection nodes and allowing track costs that exploit knowledge of the road topology and direction of travel. An example shown in Fig. 3 illustrates the first two iterations of SPAAM where the detection nodes are shown overlaid on a sequence of co-registered frames to establish the spatial and temporal context.
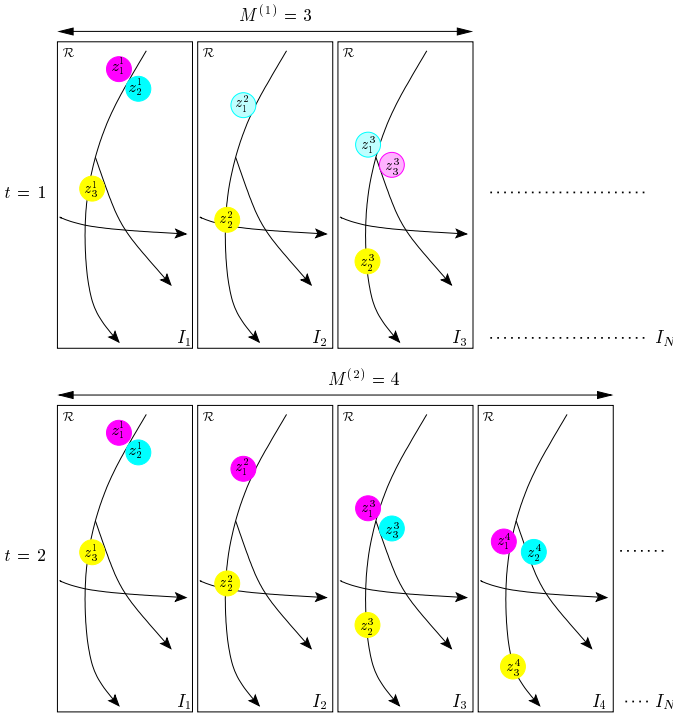

Fig. 3: An example illustrating the first two iterations of SPAAM. The variable t indicates the iteration number and for each t the rectangular boxes arranged horizontally represent successive WAMI frames, each in the common reference coordinate system $\mathcal{R}$ with the road network overlaid (as black curves with direction of travel indicated by arrows). Detected vehicle locations associated within same track are shown in a single color hue with the saturation indicating the pairwise association confidence. The first iteration in the upper row computes associations over $M^{(1)} = 3$ frame windows and hierarchically determines associations across the windows to obtain associations over the entire temporal support N to form tracks. The tracks are then assessed to evaluate a confidence probability for each pairwise association. In the second iteration shown in the lower row, pairwise associations are stochastically retained with a probability matching the estimated confidence probability and dissolved otherwise. The hierarchical process for estimating associations is then repeated but using a $M^{(2)} = 4$ frame temporal window. The process continues for further iterations (not shown).

Having summarized SPAAM at a high-level in the preceding paragraph, we next describe in detail the steps involved in the tth iteration, relying on the trellis graph formulation developed in Section II. To facilitate explanation, we use an illustrative example shown in Fig. 4 that also uses the trellis graph representation. At the start of tth iteration, an estimated set of tracks is available from the $(t-1)^{\text{th}}$ iteration as an edge induced subgraph $\mathcal{G}^{(t-1)}$ of the graph G. For each edge $\varepsilon_{k_i,k_{i+1}}^i \in E\left(\mathcal{G}^{(t-1)}\right)$, we estimate a confidence probability $P^i\left(k_i - k_{i+1}\right)$ for the pairwise association between nodes $v_{k_i}^i$ and $v_{k_{i+1}}^{i+1}$ on a path $P_k \in \mathcal{P}\left(\mathcal{G}^{(t-1)}\right)$. We consider two alternatives for estimating these confidence probabilities that are detailed in the Appendix. With the estimated probabilities we have a version of $\mathcal{G}^{(t-1)}$ where each edge is annotated with a confidence probability for the corresponding association, as indicated in the top row in the example of Fig. 4. Based on $P^i\left(k_i - k_{i+1}\right)$, the edge ($\equiv$ pairwise association) $\varepsilon_{k_i,k_{i+1}}^i$ is stochastically retained or discarded by updating the adjacency set $\mathcal{A}^{(t,0)}\left(v_{k_i}^i\right)$ for the node $v_{k_i}^i$ for the tth iteration. Specifically independent Bernoulli random variables $\eta_{k_i}^i$ are generated: with $\eta_{k_i}^i$ taking a value of 1 with probability of $P^i\left(k_i - k_{i+1}\right)$ and 0 otherwise, and the set of disassociated nodes at the next stage of the trellis and the adjacency set are, respectively, updated as

$$\mathcal{D}_{i+1}^{(t,0)} = \left\{ v_{k_{i+1}}^{i+1} | \varepsilon_{k_i,k_{i+1}}^i \in E\left(\mathcal{G}^{(t-1)}\right) \text{ and } \eta_{k_i}^i = 0 \right\}, \quad (9)$$

$$\mathcal{A}^{(t,0)}\left(v_{k_i}^i\right) = \begin{cases} v_{k_{i+1}}^{i+1} & \text{if } \eta_{k_i}^i = 1, \\ \mathcal{C}\left(v_{k_i}^i\right) & \text{if } \eta_{k_i}^i = 0, \end{cases} \quad (10)$$

where $\mathcal{C}\left(v_{k_i}^i\right)$ is a set of disassociated nodes in the $(i+1)^{\text{th}}$ frame that can be reached from $v_{k_i}^i$ given the road network, defined as

$$\mathcal{C}\left(v_j^i\right) = \begin{cases} \left\{ v_0^{i+1}, v_1^{i+1}, \dots v_{D_{i+1}}^{i+1} \right\} & j = 0, \\ \left\{ v_0^{i+1} \right\} \bigcup \\ \left\{ v_j^{i+1} | \vartheta\left(v_j^i, v_j^{i+1}\right) \leq \tau \text{ and } v_j^{i+1} \in \mathcal{D}_{i+1}^{(t,0)} \right\} & \text{otherwise,} \end{cases} \quad (11)$$

where $\vartheta\left(u_1, u_2\right)$ is the minimum distance of travel on the road network from u_1 to u_2, and τ is a threshold that is determined based on the maximum distance a vehicle can travel between successive WAMI frames. Note that, the minimum distance of travel $\vartheta\left(u_1, u_2\right)$ is determined by the shortest route on the road network between the two locations u_1 and u_2. The road network is converted to a graph [32, pp. 2-3] and the shortest route is estimated in our current implementation[4] by Dijkstra's shortest path algorithm [33, pp. 38-39]. Section S.I in the supplementary material provides additional details.

As a result of the aforementioned steps, the association between $v_{k_i}^i$ and $v_{k_{i+1}}^{i+1}$ estimated in the $(t-1)^{\text{th}}$ iteration

---

[4]Note that, the minimum distance of travel between every two points on the road network can be precomputed and stored in a look-up table for efficiency (with locations suitably quantized).

is *dis-associated*[5] with probability $\left(1 - P^i\left(k_i - k_{i+1}\right)\right)$. The proposed stochastic[6] dis-association yields adjacency sets $\mathcal{A}^{(t,0)}(.)$ that contain a significantly smaller number of nodes than those obtained by naively considering all nodes in the next frame or those obtained by considering the VDs that are spatially close. This is illustrated via an example in Fig. 5 which highlights the fact that the combined use of the aligned road network and the stochastic disassociation in SPAAM significantly reduces the number of association hypotheses to be considered compared with an approach based on spatial distances alone.

After completion of the stochastic disassociation step, the tth iteration of SPAAM continues with the hierarchical estimation of tracks. This process is schematically illustrated using the trellis graph representation starting in the second row of Fig. 4 with ℓ denoting the hierarchy level, which starts with value of 0 in the second row and increments in subsequent rows. At the ℓth level of the hierarchy, the temporal support of $\mathcal{I}$ is partitioned into disjoint temporal windows, each having $q^\ell M^{(t)}$ temporally adjacent frames. For the wth temporal window comprising frames $(w-1)q^\ell M^{(t)} + 1$ through $wq^\ell M^{(t)}$, a corresponding trellis graph $G_w^{(t,\ell)}$ is obtained using all the nodes from G over this temporal window and links between nodes defined by the adjacency set $\mathcal{A}^{(t,\ell)}(a)$ for each node a in the graph $G_w^{(t,\ell)}$. For the 0th level of the hierarchy, the adjacency sets $\mathcal{A}^{(t,0)}(a)$ for nodes $a \in V\left(\mathcal{G}_m^{(t,\ell-1)}\right)$ are as defined in (10) and for the higher levels of the hierarchy ($\ell > 0$), these sets are defined later in this section. Tracks over each temporal window w are then estimated by solving the optimization in (3) over the temporal window, i.e., by determining the optimal edge induced sub-graph $\mathcal{G}_w^{(t,\ell)}$ from $G_w^{(t,\ell)}$ as

$$\mathcal{G}_w^{(t,\ell)} = \arg \min_{\mathcal{G} \in \mathcal{F}\left(G_w^{(t,\ell)}\right)} \psi\left(\mathcal{G}\right), \qquad (12)$$

where $\psi(.)$ is as defined in (4). As shown in Fig. 4, w ranges from 1 through $N/\left(q^\ell M^{(t)}\right)$, and the temporal extent of the graph $G_w^{(t,\ell)}$ spans $s\left(G_w^{(t,\ell)}\right) = (w-1)q^\ell M^{(t)} + 1$ through $e\left(G_w^{(t,\ell)}\right) = wq^\ell M^{(t)}$. The adjacency sets for all nodes $a \in G$ are then updated for the next level of the hierarchy as we show in Fig. 4; within each temporal segment used at the current hierarchy level estimated edges are maintained, whereas across segments the results from the stochastic disassociation are used. Specifically, if i is a positive multiple of $q^\ell M^{(t)}$, we set $\mathcal{A}^{(t,\ell+1)}(v_a^i) = \mathcal{A}^{(t,0)}(v_a^i)$. Otherwise, we set $\mathcal{A}^{(t,\ell+1)}(v_a^i)$ as the singleton set $\{v_b^{i+1}\}$ where v_b^{i+1} is the node at frame $(i+1)$ estimated to associate with node a in the correspond-

---

[5]Note that, for a track that starts with trailing dummy nodes at iteration (t), we disassociate the association from the last dummy node in the trailing dummy nodes to the first non-dummy VD node of the track at iteration $(t+1)$. Similarly, for a track that ends with trailing dummy nodes at iteration (t), we disassociate the association from the last non-dummy VD node to the first dummy node in the trailing dummy nodes at iteration $(t+1)$.

[6]In our experiments, tracking results similar to those reported in this paper but with slightly larger run times, were obtained with the minor variant of SPAAM where dis-associations are obtained simply by thresholding the estimated probabilities, i.e. in (10), $\eta_{k_i}^i = 1$ if $P_i(k_i - k_{i+1}) > 0.5$ and 0 otherwise.

Algorithm 1: Stochastic progressive association across multiple frames (SPAAM)

Input : Vehicle detections $\mathcal{Z}$ for N WAMI frames
Output: Vehicle tracks $\hat{\mathcal{T}}$

1 $M^{(1)} \leftarrow M$; $V\left(\mathcal{G}^{(0)}\right) \leftarrow \bigcup_{i=1}^N \mathcal{V}_i$; Initialize $\mathcal{G}^{(0)}$;
2 **for** $t = 1$ to max iter **do**
3 $\quad$ $\ell_{max}^{(t)} \leftarrow \frac{\log N/M^{(t)}}{\log q}$;
4 $\quad$ Estimate pairwise association confidence
$\quad\quad$ $P^i\left(k_i - k_{i+1}\right)$, $\forall \varepsilon_{k_i, k_{i+1}}^i \in E\left(\mathcal{G}^{(t-1)}\right)$ (Appendix);
5 $\quad$ Update adjacency set $\mathcal{A}^{(t,0)}\left(v_{k_i}^i\right)$, $\forall v_{k_i}^i \in V\left(\mathcal{G}^{(t-1)}\right)$
$\quad\quad$ using (10);
6 $\quad$ **for** $\ell = 0$ to $\ell_{max}^{(t)}$ **do**
7 $\quad\quad$ **for** $w = 1$ to $\frac{N}{q^\ell M^{(t)}}$ **do**
8 $\quad\quad\quad$ $s_w^{(t,\ell)} \leftarrow (w-1)q^\ell M^{(t)} + 1$; $e_w^{(t,\ell)} \leftarrow wq^\ell M^{(t)}$;
$\quad\quad\quad$ **Obtain** $V\left(G_w^{(t,\ell)}\right)$ **and** $E\left(G_w^{(t,\ell)}\right)$
$\quad\quad\quad$ **to construct** $G_w^{(t,\ell)}$:
9 $\quad\quad\quad$ $V\left(G_w^{(t,\ell)}\right) \leftarrow \bigcup_{i=s_w^{(t,\ell)}}^{e_w^{(t,\ell)}} \mathcal{V}_i$;
10 $\quad\quad\quad$ **if** $\ell \neq 0$ **then**
11 $\quad\quad\quad\quad$ **for** $i = s_w^{(t,\ell)}$ to $e_w^{(t,\ell)}$ **do**
12 $\quad\quad\quad\quad\quad$ **for** $a = 1$ to D_{i+1} **do**
13 $\quad\quad\quad\quad\quad\quad$ **if** $i \mod q^{\ell-1}M^{(t)} = 0$ **then**
14 $\quad\quad\quad\quad\quad\quad\quad$ $\mathcal{A}^{(t,\ell)}\left(v_a^i\right) \leftarrow \mathcal{A}^{(t,0)}\left(v_a^i\right)$;
15 $\quad\quad\quad\quad\quad\quad$ **else**
16 $\quad\quad\quad\quad\quad\quad\quad$ $m \leftarrow \left\lfloor i/\left(q^{\ell-1}M^{(t)}\right)\right\rfloor$;
17 $\quad\quad\quad\quad\quad\quad\quad$ $\mathcal{A}^{(t,\ell)}\left(v_a^i\right) \leftarrow \{v_b^{i+1}|\varepsilon_{a,b}^i \in$
$\quad\quad\quad\quad\quad\quad\quad\quad E\left(\mathcal{G}_m^{(t,\ell-1)}\right)\}$;
18 $\quad\quad\quad\quad\quad$ **end**
19 $\quad\quad\quad\quad$ **end**
20 $\quad\quad\quad$ **end**
21 $\quad\quad\quad$ **end**
22 $\quad\quad\quad$ $E\left(G_w^{(t,\ell)}\right) \leftarrow$
$\quad\quad\quad\quad \left\{\varepsilon_{a,b}^i \middle| v_b^{i+1} \in \mathcal{A}^{(t,\ell)}\left(v_a^i\right), v_a^i, v_b^{i+1} \in V\left(G_w^{(t,\ell)}\right)\right\}$;
23 $\quad\quad\quad$ Estimate $\mathcal{G}_w^{(t,\ell)}$ from $G_w^{(t,\ell)}$ using (12);
24 $\quad\quad$ **end**
25 $\quad$ **end**
26 $\quad$ $\mathcal{G}^{(t)} \leftarrow \mathcal{G}_1^{(t,\ell_{max}^{(t)})}$; $M^{(t+1)} \leftarrow M^{(t)} + 1$;
27 **end**
28 $\hat{\mathcal{T}} \leftarrow \mathcal{P}\left(\mathcal{G}^{(t)}\right)$;

---

ing temporal segment at level ℓ of the hierarchy; formally $\varepsilon_{a,b}^i \in E\left(\mathcal{G}_m^{(t,\ell)}\right)$ where $m = \left\lfloor i/\left(q^\ell M^{(t)}\right)\right\rfloor$.

The edge induced sub-graph $\mathcal{G}_1^{(t,\ell_{max})}$ estimated at the final final hierarchy level $\ell_{max}^{(t)} = \log N/\left(M^{(t)}\log q\right)$ becomes the estimated edge induced subgraph $\mathcal{G}^{(t)}$ at the end of the tth iteration defining the estimates of the optimal paths and associations upon completion of the tth iteration. The edge induced subgraph $\mathcal{G}^{(t)}$ then forms the starting point for the $(t+1)^{\text{th}}$ iteration. The overall SPAAM algorithm is summarized[7] in Algorithm 1.

---

[7]The decomposed BIP problem (line 23 of Algorithm) can be solved using a variety of integer linear programming solvers such as [34]–[38].
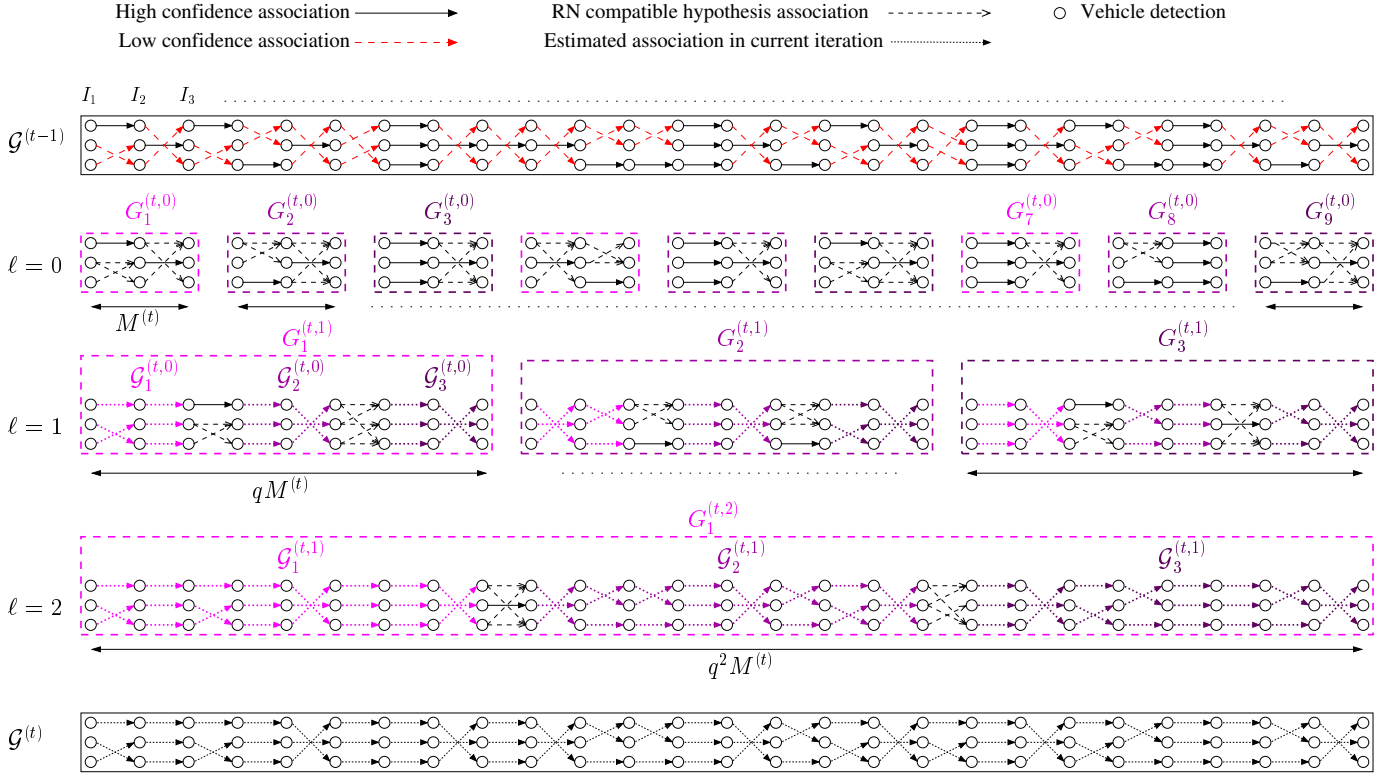
Fig. 4: Hierarchical estimation of tracks in the tth iteration of SPAAM. At the start of the tth iteration, a confidence probability is estimated for every pairwise association (edge) in $\mathcal{G}^{(t-1)}$. Associations from $\mathcal{G}^{(t-1)}$ are retained or dissolved based on the estimated confidence probabilities. Remaining associations are then hierarchically estimated in a common framework aimed at minimizing (3). At the $\ell = 0^{\text{th}}$ level of the hierarchy, the trellis graph is partitioned into windows of length $M^{(t)}$ and for each trellis graph $G_w^{(t,\ell)}$ over the wth temporal window associations are estimated with the objective of minimizing (3). Subsequent levels of the hierarchy ($\ell > 0$) estimate associations for temporal windows formed by groups of q adjacent windows from hierarchy level ($\ell - 1$) using the same methodology: preserving retained associations and estimating remaining associations with the objective of minimizing (3). Upon completion of the hierarchical estimation process for the tth iteration of SPAAM, the set of retained and estimated associations define the edge induced subgraph $\mathcal{G}^{(t)}$ that summarizes the estimated tracks and forms the input for the next iteration.

IV. EXPERIMENTAL RESULTS

To evaluate tracking performance, we use three test sequences with distinct characteristics chosen from two WAMI data sets: (a) CorvusEye dataset that is recorded using the CorvusEye 1500 Wide-Area Airborne System [39] over the Rochester, NY region, and (b) Wright-Patterson Air Force Base (WPAFB) 2009 dataset [40], which was recorded over the WPAFB, OH region. Fig. 6 depicts a frame from each sequence of the three sequences that we label as "Seq1", "Seq2" and "Seq3". "Seq1" uses a region from the CorvusEye dataset that contains forked one-way roads with different directions and also several occluders (bridges, trees, etc.). "Seq2" and "Seq3" correspond to two different regions from the WPAFB dataset. Both "Seq2" and "Seq3" cover regions that are free from occluders, however, "Seq2" contains only two-way roads, "Seq3" has only one-way roads. All test sequences comprise 60 ground truth labeled frames[8]. For the

vector road map, we use OpenStreetMap (OSM) [41], which provides each road in the road network in a vector format along with properties of each road such as type (highway, residential, etc), one or two-way traffic, number of lanes, etc. We use the method in [26] for co-registering the WAMI frames to the vector road map. We obtain the VDs via the background subtraction method used in [13]. We set $M^{(1)} = 3$ in all experiments and we use the Gurobi optimizer [34] for solving (12). The Gurobi optimizer relaxes the integer restrictions in the binary constraints in (1) and uses a linear-programming based branch-and-bound algorithm for obtaining a solution[9]. For all experiments, the parameters in (5) are set to empirically determined values of $\sigma_m = 12.5$, $\sigma_d = 0.02$, $\sigma_\theta = 100$, $\sigma_g = 2.8$, and $W = 2$.

Several methods are benchmarked. Two instances of the proposed method are considered: **SPAAM-M** and **SPAAM-EM**, which, respectively, estimate the confidence of pairwise associations using the marginalization in (13) or the EM approach in (16). We compare the proposed SPAAM

[8]The detection locations that form the inputs to the algorithm, along with the aligned road network, and labeled ground truth tracks of the three sequences are available at http://www.ece.rochester.edu/~gsharma/VisDataAnalGeoSpat/

[9]More information about the algorithm used by the Gurobi optimizer can be found at http://www.gurobi.com/resources/getting-started/mip-basics

(a) Spatial distance   (b) Road network   (c) Stochastic dis-association + road network
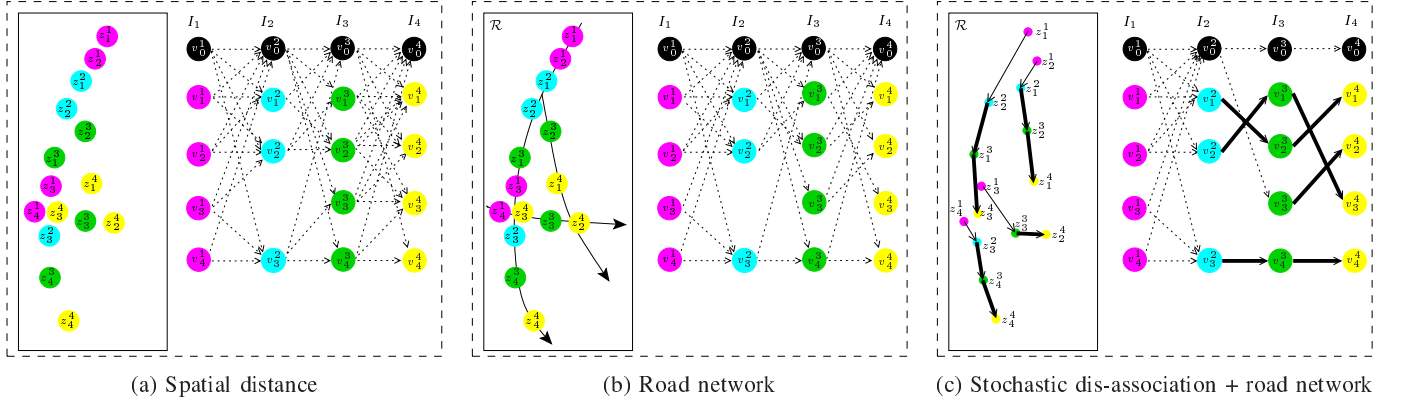
Fig. 5: An example comparing alternative strategies for generating association hypotheses for a four frame WAMI sequence I_1, I_2, I_3, and I_4 at the first hierarchy level ($\ell = 0$) in the tth iteration. Distinct colors identify the VDs detected in a single WAMI frame and hypothesis associations are represented by directed edges in a corresponding trellis graph. Sub-figure (a) illustrates a situation where the adjacency set for each node and the graph G are formed without using the road network. In this situation, as a specific example, note that there are three nodes (plus the dummy node) that are spatially close with v_3^3, so the adjacency set $\mathcal{A}^{(t,0)}\left(v_3^3\right)$ determined purely based on distance contains 4 nodes. Sub-figure (b) illustrates the situation where the adjacency set for each node and the graph G are formed by using the aligned road network. The graph is sparser and the adjacency sets for each node are smaller than in Fig. 5 (a). Specifically, for example, $\mathcal{A}^{(t,0)}\left(v_3^3\right)$ will contains only two nodes because v_2^4 is the only node within a reasonable road travel distance from v_3^3. Sub-figure (c) illustrates that the proposed stochastic disassociation approach where the confidence of estimated pairwise associations is represented by widths of the directed lines. The proposed approach further reduces the adjacency set for each node, specifically, in this example the association between v_3^3 and v_2^4 is maintained from the previous iteration and therefore, $\mathcal{A}^{(t,0)}\left(v_3^3\right)$ contains only one node.

instances with the multi-data association, iterative conditional modes like method (**MDA-ICM**) of [19]. The MDA-ICM method does not require per association independent additive costs for tracks and is used with the same cost function as SPAAM (defined in (5)) allowing for a fair comparison of the techniques without differences induced by the underlying cost function. The methods are also compared against an online tracking method **OT** proposed in [12] that considers a pre-aligned road network when estimating the frame-to-frame VDs associations. The tracks estimated by the online tracking method also serve as the initialization for all of the iterative methods, specifically, the different instances of SPAAM and for MDA-ICM.

We quantify tracking performance with reference to labeled ground-truth tracks by the widely adopted measures defined in [42] which are (1) the total number of ID switches ($IDS\downarrow$)[10] for estimated tracks compared to the ground truth, (2) the number of mostly tracked ($MT\uparrow$) vehicles, i.e., the vehicles for which estimated tracks include over 80% of the detections in the ground truth, (3) the number of mostly lost ($ML\downarrow$), i.e., the vehicles for which estimated tracks include under 20% of the detections in the ground truth, (4) the number of partially tracked ($PT\downarrow$), i.e., the vehicles for which estimated tracks include below 80% and over 20% of the detections in the ground truth, and (5) the number of track fragments ($Frag\downarrow$), where a fragment is defined as a part of vehicle track with length less than 80% of the length of the corresponding ground truth track. In addition to these measures, we report the Multiple Object Tracking Accuracy ($MOTA\uparrow$) defined in [43] (with $c_s = 1$). Table I enumerates these metrics for the

Seq	Method	MT↑	ML↓	PT	IDS↓	Frag↓	MOTA↑
1	OT	73	11	55	292	384	0.854994
	MDA-ICM	81	15	43	164	211	0.91953
	SPAAM-M	85	14	40	110	176	0.946249
	SPAAM-EM	80	16	43	**82**	**144**	**0.959503**
2	OT	94	2	6	311	360	0.884093
	MDA-ICM	95	3	4	257	290	0.902187
	SPAAM-M	97	3	2	101	128	0.961854
	SPAAM-EM	97	3	2	**60**	**83**	**0.977017**
3	OT	53	0	1	112	138	0.932051
	MDA-ICM	51	1	2	86	101	0.944871
	SPAAM-M	54	0	0	40	59	0.974878
	SPAAM-EM	54	0	0	**27**	**39**	**0.983806**

TABLE I: Comparison of tracking performance metrics for four different tracking algorithms on the three different WAMI sequences (datasets). The metrics include mostly tracked (MT), mostly lost (ML), partly tracked (PT), ID switches (IDS), fragments (Frag), and multiple object tracking object accuracy (MOTA). See text for definitions of the metrics and methods. The annotations ↑ and ↓ next to a metric in the table indicate whether a higher or lower value is better, respectively.

tracks estimated for the three WAMI datasets using the four methods: OT method in [12], the two proposed instances of SPAAM (SPAAM-M and SPAAM-EM), and the MDA-ICM method [19]. The results in the table indicate that the proposed SPAAM instances and the MDA-ICM method both improve upon the OT method that serves as the initialization for both these methods. However, the improvement offered by the proposed SPAAM instances is much larger than that for the MDA-ICM method. Overall, both the proposed SPAAM instances (SPAAM-M and SPAAM-EM) perform significantly better than the MDA-ICM method. Additional results that

---

[10]↑ and ↓ indicate that higher and lower is better, respectively.

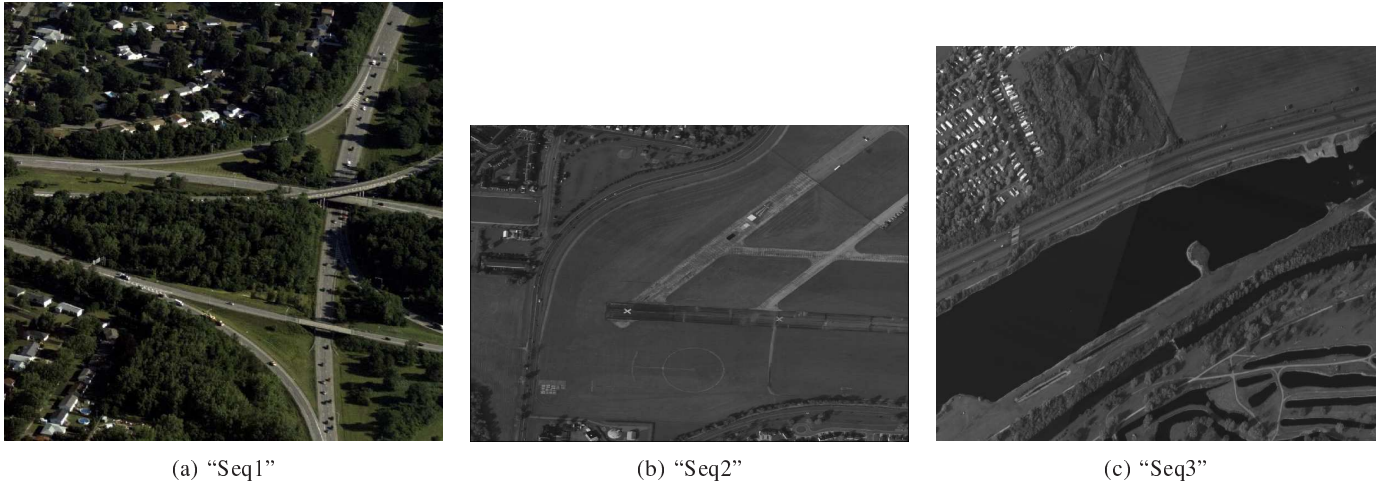(a) "Seq1"   (b) "Seq2"   (c) "Seq3"

Fig. 6: Test sequences used in our evaluation. Each test sequence contains 60 frames and we show the first frame from "Seq1", "Seq2", and "Seq3", in (a), (b), and (c) respectively. All roads in "Seq1" and "Seq3" are one-way roads, while roads in "Seq2" are two-way roads.

Seq	SPAAM-EM		SPAAM-M		MDA-ICM	
	Iters	Time	Iters	Time	Iters	Time
1	6	0.1490	6	0.1479	23	26.8306
2	6	0.2003	6	0.2142	13	6.1653
3	6	0.0833	6	0.0235	22	3.4917

TABLE II: Execution time (in hours) for the two proposed SPAAM instances (SPAAM-M and SPAAM-EM) and for the MDA-ICM method along with the number of corresponding iterations (Iters).

illustrate the effect of each term in the cost (5) on the overall performance of SPAAM-EM is presented in Section S.II in the supplementary material.

Table II lists the execution times for the two proposed SPAAM instances (SPAAM-M and SPAAM-EM) and for the MDA-ICM method along with the corresponding number of iterations. All algorithms are executed on a computer with an Intel® Core™ i7-6700HQ CPU with 8 cores, operating at 2.60 GHz, having 12 GB of main memory, and running the Ubuntu 16.04 LTS operating system. From the table it can be seen that the proposed SPAAM instances have much lower execution times. Thus the proposed SPAAM approach offers performance gains in both tracking accuracy and in execution time.

V. DISCUSSION

The performance of the OT method is hampered by the fact that it minimizes a cost function that is a sum of independent costs for pairwise associations between adjacent frames. In situations like WAMI vehicle tracking, where only vehicle detection locations are available, such cost functions are fundamentally limited because they do not allow the use of a larger temporal context for assessing plausibility of (postulated) vehicle movements. As a result, although OT method minimizes its cost function via the estimation of associations from frame to frame in temporal sequence, the limitations of the cost function make it prone to tracking errors. The MDA-ICM and

proposed SPAAM methods, use a cost function that is better suited for WAMI vehicle tracking by using a larger temporal context to assess vehicular movements. This allows them to improve upon the performance of the OT method, even though the methods are not able to guarantee that a global minimum of the cost function is achieved. To further understand the differences between the MDA-ICM and the proposed SPAAM approach, it is instructive to compare the evolution of the (common) cost function for these methods with the progression of iterations. Figure 7 shows the total cost for tracks estimated using the proposed SPAAM instances (SPAAM-M and SPAAM-EM) and the MDA-ICM method as a function of the iteration number for these methods. In each step of its iterative process, the MDA-ICM method re-estimates all of the associations across temporally adjacent frames to improve the global cost, while keeping all other associations fixed. This approach guarantees that the iterations improve monotonically assuring convergence of the MDA-ICM approach to a local minimum, a property it inherits from the iterative conditional modes (ICM) [44] approach on which it is based. While the approach ensures computational tractability, it cannot avoid getting trapped in poor local minima in the vicinity of the initialization. On the other hand, via the stochastic disassociation and progressive enlargement of temporal context, the SPAAM approach revisits (predominantly) low confidence associations collectively with the benefit of additional information available from the enlarged context. Thus, even though SPAAM does not guarantee monotonic convergence, it is better at avoiding local minima and, as demonstrated in the results, typically offers a significant improvement in tracking accuracy.

To highlight the benefit of stochastic dis-association and the progressive enlargement of the temporal window sizes in SPAAM, we consider **SPAAM–**, a deliberately de-featured version of proposed approach that estimates all of the associations in single iteration using a fixed temporal window size. The SPAAM– approach is obtained in Algorithm 1 by setting *max iterations* = 1 and $E\left(\mathcal{G}^{(0)}\right) = \phi$. Fig. 8 compares the execution times required for estimation of tracks for "Seq1"
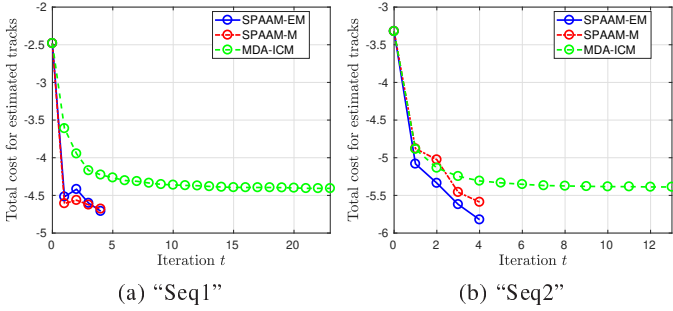
(a) "Seq1"   (b) "Seq2"

Fig. 7: The total cost of the estimated tracks with progression of iterations for the proposed SPAAM instances (SPAAM-M and SPAAM-EM) and for the MDA-ICM method in [19] for: (a) "Seq1" and (b) "Seq2". The methods use a common cost function and a common initialization. The MDA-ICM technique demonstrates monotonic improvement of the cost with iterations, whereas SPAAM-M and SPAAM-EM do not exhibit monotonic improvement. However, with the progression of iterations, SPAAM-M and SPAAM-EM achieve a lower total cost than MDA-ICM, yielding a better set of estimated tracks.
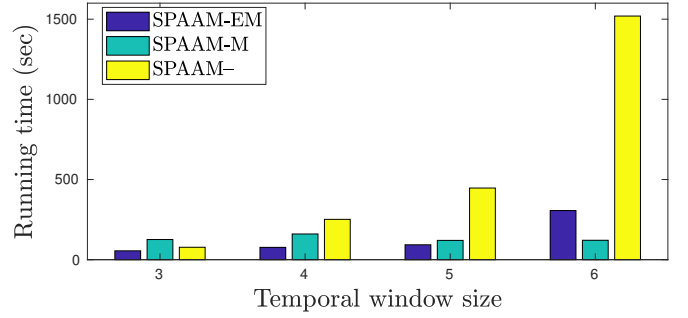

Fig. 8: Execution time (in seconds) for different temporal window sizes for SPAAM-M, SPAAM-EM, and the de-featured variant SPAAM– that estimates all associations afresh. "Seq1" was used for obtaining these execution times. The figure illustrates the key advantages of the SPAAM-M and SPAAM-EM approaches: with progressive enlargement of the temporal window size the computation increases only modestly, whereas the execution time increases exponentially with increasing window size for SPAAM–. The hardware configuration used for these experiments is identical to that specified in Table II.

using the proposed SPAAM-M, SPAAM-EM, and the de-featured variant SPAAM– with different temporal window sizes[11]. The figure shows that the execution time for the SPAAM– increases exponentially with the increase in the size of the temporal window used for estimating the associations. On the other hand, the SPAAM-M and SPAAM-EM approaches experience only a small increase in the execution time with progressive enlargement of the temporal window sizes because, with high probability, they maintain pairwise associations already estimated in previous iteration with high confidence in the previous iteration and revisit only the low confidence associations. Thus, the number of association hypotheses has a limited increase with the enlargement of the temporal window size, which is reflected as a corresponding small increase in the execution time seen in the figure.

As motivated earlier, the tracking by detection framework used in this paper is well suited to the WAMI setting because of the limited spatial detail for individual vehicles in WAMI frames. However, limitations of the tracking by detection framework also carry over to the proposed approach. For instance, "split and merge" situations where, over the observed time course, multiple vehicles are either detected individually or as a single merged detection pose a challenge for tracking by detection and thereby for the proposed approach. We note that for "short duration merges" where the detections are separate, merge for a few frames, and then separate, the proposed approach can treat the merged detections as one vehicle and the other as "occluded". The cost functions that assess the regularity of vehicular motion over a multi-frame temporal window allow SPAAM to link tracks across the occlusion window.

Finally, we note that although this paper focused specifically on the problem of vehicle tracking in WAMI, the SPAAM approach is more general and may offer a useful framework for other trellis graph association problems where: (a) it is beneficial to use cost functions that use a larger temporal context that disallows a pairwise independent cost decomposition and (b) one can assess confidence of associations. In this setting, we anticipate that SPAAM can offer an advantage over alternative stochastic optimization approaches such as simulated annealing or genetic algorithms because it explicitly exploits the trellis graph structure. Exploration of such alternatives is beyond the scope of the present paper but would make interesting future work.

VI. CONCLUSION

The stochastic progressive association across multiple frames (SPAAM) framework proposed in this paper provides a novel, computationally-efficient, iterative, and accurate approach for estimating vehicular tracks in WAMI. Specifically, SPAAM allows the use of effective cost functions that assess the regularity of vehicular motion over a multi-frame window while simultaneously maintaining computational tractability. These dual objectives are achieved via a strategy of estimating tracks hierarchically over progressively increasing window lengths. As the window is enlarged, high confidence estimated associations are predominantly maintained and low confidence associations are predominantly dissolved, and tracks are re-estimated with the larger temporal context. The stochastic approach to maintaining/dissolving associations reduces the computation required for larger windows while allowing the tracks to be improved through the incorporation of additional information provided by the larger context. The combination of SPAAM with information provided by a co-registered road network allows the proposed method to tackle the challenges of scale and limited spatial detail in WAMI tracking. Results obtained over three test sequences that represent different tracking scenarios show a significant performance improve-

---

[11]For SPAAM-M and SPAAM-EM these are the corresponding iteration execution time.

ment for the proposed approach compared with three state of the art alternative methods.

APPENDIX

Confidence for pairwise associations is estimated via two alternative probabilistic modeling approaches. The first, uses an energy model to transform costs for tracks into probabilities, followed by approximate marginalization to estimate the probability of an association. The second approach uses a two component mixture model for motion features. Parameters for the model are obtained using semi-supervised expectation maximization (EM) and the association probability is estimated as the posterior probability of the corresponding mixture component.

Energy model based marginalization for pairwise association confidence estimation

An energy model [45] can be used to coherently extend the costs we already use into a corresponding probabilistic model. Specifically, the probability of an edge induced sub-graph $\mathcal{G} \in \mathcal{F}(G)$ can be modeled as $P(\mathcal{G}) = \frac{1}{Z_G} e^{-\psi(\mathcal{G})}$, where $Z_G = \sum_{\mathcal{G} \in \mathcal{F}(G)} e^{-\psi(\mathcal{G})}$ is partition function that ensures probabilities sum to 1. The minimization in (3) is then equivalent to a maximum a posterior (MAP), or minimum energy, estimate. With the probabilistic model, the confidence $\bar{P}^i(a-b)$ of a pairwise association that associates v_a^i with v_b^{i+1} can be estimated as the marginal probability of inclusion of the edge $\varepsilon_{a,b}^i$ in the sub-graph. Formally, $\bar{P}^i(a-b) = \sum_{\substack{\mathcal{G} \in \mathcal{F}(G), \\ \varepsilon_{a,b}^i \in E(\mathcal{G})}} P(\mathcal{G})$.

The above marginalization is intractable because of the dependency introduced by the feasibility constraints. Therefore, we approximate $\bar{P}^i(a-b)$ by

$$P^i(a-b) = \frac{1}{Z_a^i} \sum_{\substack{P_k \in \mathcal{P}(G), \\ v_a^i, v_b^{i+1} \in P_k}} e^{-C(P_k)}, \qquad (13)$$

where $Z_a^i = \sum_{v_m^{i+1} \in \mathcal{A}(v_a^i)} \sum_{\substack{P_k \in \mathcal{P}(G), \\ v_a^i, v_m^{i+1} \in P_k}} e^{-C(P_k)}$ is a normalization

constant. The above approximation is reasonable for the following reasons that are apparent from (13). First, v_a^i must be associated with only one of the VDs in $\mathcal{A}(v_a^i)$, and therefore, $\sum_{v_m^{i+1} \in \mathcal{A}(v_a^i)} P^i(a-m) = 1$. Second, associating v_a^i with v_b^i is more probable than associating with v_c^i if the sum of the costs of the tracks that contain the pairwise association between v_a^i and v_b^i is smaller than the tracks that contain the pairwise association between v_a^i and v_c^i.

Semi-supervised EM based mixture model

To estimate the confidence of a pairwise association $\varepsilon_{a,b}^i$ that associates v_a^i with v_b^{i+1}, we first estimate temporally local features $\mathbf{f}_{a,b}^i = \left[\Gamma_m\left(\varepsilon_{a,b}^i\right), \Gamma_\theta\left(\varepsilon_{a,b}^i\right), R_d\left(\varepsilon_{a,b}^i\right), R_\theta\left(\varepsilon_{a,b}^i\right) \right]$ for the pairwise association, then we model the distribution of $\mathbf{f}_{a,b}^i$ as a two component Gaussian mixture model (GMM) [45]. One mixture component corresponds to the case when v_a^i *associates* with v_b^{i+1} (denoted as $v_a^i \rightarrow v_b^{i+1}$), while the other component corresponds to the case when v_a^i *does not associate* with v_b^{i+1}. To indicate which mixture component is responsible for the pairwise association, we associate the pairwise association with a binary latent variable $c_{a,b}^i$, where

$$c_{a,b}^i = \begin{cases} 1 & \text{if } v_a^i \rightarrow v_b^{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

The confidence of the pairwise association is then approximated by the posterior probability that $c_{a,b}^i = 1$ given the computed features $\mathbf{f}_{a,b}^i$.

Specifically, the two GMM components are parametrized by the corresponding means $\boldsymbol{\mu}_r$ and covariances $\boldsymbol{\Sigma}_r$ of the features, where $r \in \{0, 1\}$. Also, $c_{a,b}^i$ is assumed to be Bernoulli with parameter $\gamma = p\left(c_{a,b}^i = 1\right)$. Thus, the GMM is parametrized by $\boldsymbol{\theta} = \{\gamma, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1\}$. The likelihood of $\mathbf{f}_{a,b}^i$ is

$$p\left(\mathbf{f}_{a,b}^i \mid \boldsymbol{\theta}\right) = \sum_{r=0}^1 p\left(\mathbf{f}_{a,b}^i, c_{a,b}^i = r \mid \boldsymbol{\theta}\right), \qquad (14)$$

where

$$\begin{aligned} p\left(\mathbf{f}_{a,b}^i, c_{a,b}^i = r \mid \boldsymbol{\theta}\right) &= p\left(c_{a,b}^i = r \mid \boldsymbol{\theta}\right) p\left(\mathbf{f}_{a,b}^i \mid c_{a,b}^i = r, \boldsymbol{\theta}\right), \\ &= \gamma^r \left(1-\gamma\right)^{1-r} \mathcal{N}\left(\mathbf{f}_{a,b}^i \mid \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r\right). \end{aligned} \qquad (15)$$

The confidence of the pairwise association is then computed as

$$\begin{aligned} P^i(a-b) &= p\left(c_{a,b}^i = 1 \,\middle|\, \mathbf{f}_{a,b}^i, \hat{\boldsymbol{\theta}}\right), \\ &= \frac{p\left(\mathbf{f}_{a,b}^i, c_{a,b}^i = 1 \,\middle|\, \hat{\boldsymbol{\theta}}\right)}{\sum_{r=0}^1 p\left(\mathbf{f}_{a,b}^i, c_{a,b}^i = r \,\middle|\, \hat{\boldsymbol{\theta}}\right)}, \end{aligned} \qquad (16)$$

where $\hat{\boldsymbol{\theta}} = \left\{\hat{\gamma}, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_0, \hat{\boldsymbol{\Sigma}}_1\right\}$ are the estimated parameters of the GMM. We estimate $\hat{\boldsymbol{\theta}}$ using the EM framework [46] through a semi-supervised approach. The semi-supervised approach exploits available ground truth training data (for example from [40]). Let $\tilde{\mathcal{G}}$ be a ground truth trellis graph[12] that is constructed from the ground truth tracks such that

---

[12]Note that the ground truth trellis graph is constructed using ground truth tracks in different dataset, i.e., $V(\mathcal{G}) \neq V(\tilde{\mathcal{G}})$.

every ground truth track represents a path in $\tilde{\mathcal{G}}$. Then, the EM framework estimates $\hat{\boldsymbol{\theta}}$ by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{\varepsilon_{a,b}^i \in E(\tilde{\mathcal{G}})} p\left(\varepsilon_{a,b}^i, c_{a,b}^i = 1 \,|\, \boldsymbol{\theta}\right) \times$$
$$\prod_{\varepsilon_{a,b}^i \in E(\mathcal{G})} p\left(\mathbf{f}_{a,b}^i \,|\, \boldsymbol{\theta}\right). \qquad (17)$$

under the assumption that the pairwise associations are conditionally independent[13].

References

[1] A. Elliethy and G. Sharma, "Vehicle tracking in wide area motion imagery : A facility location motivated combinatorial approach," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, March 5-9, 2017, New Orleans, Mar. 2017, pp. 1932–1936.

[2] J. Hawkes, *Aerial: The Art of Photography From the Sky*. Hove, UK: Rotovision, 2003.

[3] K. Palaniappan, R. M. Rao, and G. Seetharaman, "Wide-area persistent airborne video: Architecture and challenges," in *Distributed Video Sensor Networks*. London, UK: Springer, 2011, pp. 349–371.

[4] E. Blasch, G. Seetharaman, S. Suddarth, K. Palaniappan, G. Chen, H. Ling, and A. Basharat, "Summary of methods in wide-area motion imagery (WAMI)," in *Proc. SPIE*, vol. 9089, 2014, pp. 90 890C–90 890C–10.

[5] R. Porter, A. Fraser, and D. Hush, "Wide-area motion imagery," *IEEE Sig. Proc. Mag.*, vol. 27, no. 5, pp. 56–65, Sept 2010.

[6] W. Luo, J. Xing, X. Zhang, X. Zhao, and T.-K. Kim, "Multiple object tracking: A literature review," *arXiv preprint arXiv:1409.7618*, 2014.

[7] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2008, pp. 1–8.

[8] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 33, no. 9, pp. 1806–1819, 2011.

[9] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, 2011, pp. 1201–1208.

[10] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2015, pp. 5537–5545.

[11] P. Lenz, A. Geiger, and R. Urtasun, "FollowMe: Efficient online min-cost flow tracking with bounded memory and computation," in *IEEE Intl. Conf. Comp. Vision.*, Dec 2015, pp. 4364–4372.

[12] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2010, pp. 679–684.

[13] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *Proc. European Conf. Computer Vision*, 2010, pp. 186–199.

[14] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. European Conf. Computer Vision*. Springer, 2008, pp. 788–801.

[15] M. Hofmann, M. Haag, and G. Rigoll, "Unified hierarchical multi-object tracking using global data association," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog. Workshops*, Jan 2013, pp. 22–28.

[16] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2013, pp. 1830–1837.

[17] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association by online target-specific metric learning and coherent dynamics estimation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. PP, no. 99, pp. 1–1, 2016.

[18] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] R. T. Collins, "Multitarget data association with higher-order motion models," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, 2012, pp. 1744–1751.

[20] X. Shi, H. Ling, J. Xing, and W. Hu, "Multi-target tracking by rank-1 tensor approximation," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2013, pp. 2387–2394.

[21] A. A. Butt and R. T. Collins, "Multiple target tracking using frame triplets," in *Asian Conf. on Comp. Vision*. Springer, 2012, pp. 163–176.

[22] ——, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, June 2013.

[23] C. R. Reeves, *Modern heuristic techniques for combinatorial problems*. John Wiley & Sons, Inc., 1993.

[24] N. Ukita and A. Okada, "High-order framewise smoothness-constrained globally-optimal tracking," *Computer Vision and Image Understanding*, vol. 153, pp. 130 – 142, 2016, special issue on Visual Tracking.

[25] J. P. Snyder, *Map projections–A working manual*. US Government Printing Office, 1987, vol. 1395.

[26] A. Elliethy and G. Sharma, "Automatic registration of vector road maps with wide area motion imagery by exploiting vehicle detections," *IEEE Trans. Image Proc.*, vol. 25, no. 11, pp. 5304 – 5315, Nov. 2016.

[27] ——, "Accelerated parametric chamfer alignment using a parallel, pipelined GPU realization," *Journal of Real-Time Image Processing*, Feb 2017.

[28] ——, "Vector road map registration to oblique wide area motion imagery by exploiting vehicles movements," in *IS&T Electronic Imaging: Video Surveillance and Transportation Imaging Applications*, San Francisco, California, 2016, pp. VSTIA–520.1–8. [Online]. Available: http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000003/art00008

[29] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Mineola, NY: Dover, 1998.

[30] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Proc. European Conf. Computer Vision*. Springer, 2012, pp. 343–356.

[31] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, 2015, pp. 4091–4099.

[32] R. Balakrishnan and K. Ranganathan, *A Textbook of Graph Theory*, ser. Universitext (Berlin. Print). Springer New York, 2012.

[33] D. Medhi and K. Ramasamy, *Network Routing: Algorithms, Protocols, and Architectures*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007.

[34] "Gurobi optimizer," accessed Mar. 2016. [Online]. Available: http://www.gurobi.com/products/gurobi-optimizer

[35] "GNU Linear Programming Kit (GLPK)," accessed July 2017. [Online]. Available: https://www.gnu.org/software/glpk/

[36] "Solving Constraint Integer Programs (SCIP)," accessed July 2017. [Online]. Available: http://scip.zib.de/

[37] "CPLEX optimizer," accessed July 2017. [Online]. Available: https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

[38] "Coin-or branch and cut (Cbc)," accessed July 2017. [Online]. Available: https://projects.coin-or.org/Cbc

[39] "CorvusEye™1500 Data Sheet," https://www.harris.com/sites/default/files/downloads/solutions/55533_corvuseyebrochure_eng_final_2_web.pdf, accessed Mar. 2018.

[40] "AFRL WPAFB 2009 data set," https://www.sdms.afrl.af.mil, accessed Mar. 2016.

[41] "OpenStreetMap," http://www.openstreetmap.org, accessed Mar. 2016.

[42] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, vol. 1, 2006, pp. 951–958.

[43] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 31, no. 2, pp. 319–336, Feb 2009.

[44] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.

[45] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

---

[13]Note that $\hat{\boldsymbol{\theta}}$ can be also estimated through an unsupervised approach, for which $E\left(\tilde{\mathcal{G}}\right) = \phi$. To identify the mixture component that is corresponds to $v_a^i \to v_b^{i+1}$ case, the estimated covariances of the two Gaussian components are compared and the component that has smaller parameter variance is selected.

Ahmed Elliethy (S'15) received the B.Sc. degree (excellent with honors) in computer engineering and the M.Sc. degree in electrical engineering from the Military Technical College, Cairo, Egypt, in 2003, and 2010, respectively. He received the PhD. degree in electrical engineering from University of Rochester, USA, in 2017. He is currently working for the Military Technical College as an assistant processor in the computer department. His research interests include computer vision and security, specifically, multiple object tracking, optical flow, and media forensics.

Gaurav Sharma (S'88–M'96–SM'00–F'13) is a professor at the University of Rochester in the Department of Electrical and Computer Engineering, in the Department of Computer Science and in the Department of Biostatistics and Computational Biology. From 2008–2010, he served as the Director for the Center for Emerging and Innovative Sciences (CEIS), a New York state funded center for promoting joint university-industry research and technology development, which is housed at the University of Rochester. He received the BE degree in Electronics and Communication Engineering from Indian Institute of Technology Roorkee (formerly Univ. of Roorkee), India in 1990; the ME degree in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, India in 1992; and the MS degree in Applied Mathematics and PhD degree in Electrical and Computer Engineering from North Carolina State University, Raleigh in 1995 and 1996, respectively. From Aug. 1996 through Aug. 2003, he was with Xerox Research and Technology, in Webster, NY, initially as a Member of Research Staff and subsequently at the position of Principal Scientist.

Dr. Sharma's research interests include data analytics, image processing and computer vision, color science and imaging, media security, and distributed signal processing. He is the editor of the "Color Imaging Handbook", published by CRC press in 2003. He is a fellow of the IEEE, of SPIE, and of the Society of Imaging Science and Technology (IS&T) and a member of Sigma Xi. He served as a Technical Program Chair for the 2016 and 2012 editions of the IEEE International Conference on Image Processing (ICIP), as the Symposium Chair for the 2013 SPIE/IS&T Electronic Imaging symposium, as the 2010-2011 Chair IEEE Signal Processing Society's Image Video and Multi-dimensional Signal Processing (IVMSP) technical committee, the 2007 chair for the Rochester section of the IEEE and the 2003 chair for the Rochester chapter of the IEEE Signal Processing Society. Dr. Sharma is the Editor-in-Chief (EIC) for the IEEE Transactions on Image Processing (TIP). From 2011 through 2015, he served as the EIC for the Journal of Electronic Imaging (JEI) and in the past has served as an associate editor for JEI, IEEE TIP, and the IEEE Transactions on Information Forensics and Security.