On Pairwise Costs for Network Flow Multi-Object Tracking

Visesh Chari[1], Simon Lacoste-Julien[2], Ivan Laptev[1], Josef Sivic[1]
[1]WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris, France
[2]SIERRA project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548, Paris, France

Multi-object tracking has been recently approached with the min-cost network flow optimization techniques. Such methods simultaneously resolve multiple object tracks in a video and enable modeling of dependencies among tracks. Min-cost network flow methods also fit well within the "tracking-by-detection" paradigm where object trajectories are obtained by connecting per-frame outputs of an object detector. Object detectors, however, often fail due to occlusions and clutter in the video. To cope with such situations, we propose to add pairwise costs to the min-cost network flow framework. While integer solutions to such a problem become NP-hard, we design a convex relaxation solution with an efficient rounding heuristic which empirically gives certificates of small suboptimality. We evaluate two particular types of pairwise costs and demonstrate improvements over recent tracking methods in real-world video sequences.

Given a video with objects in motion, the goal is to simultaneously track K moving objects in a "detect-and-track" framework [2]. The input to the approach is two-fold. First a set of candidate object locations is assumed to be given, provided, for example, as output of an object detector. Henceforth we refer to these locations as *detections*. The approach also requires a measure of correspondence between detections across video frames. This could be obtained for example from optical flow, or using some other form of correspondence. Based on these inputs, the tracking problem is setup as a joint optimization problem of simultaneously selecting detections of objects and connections between them across video frames. Such a problem can be modeled through a MAP objective [2] with specific constraints encoding the structure of the tracks. The MAP optimization problem can be cast as the following integer linear program (ILP):

$$\begin{aligned} \min_{\mathbf{x}} \quad & \textstyle\sum_i c_i x_i + \sum_{ij \in E} c_{ij} x_{ij} \qquad (1) \\ \text{s.t.} \quad & \left.\begin{array}{c} 0 \le x_i \le 1 \,, 0 \le x_{ij} \le 1 \\ \displaystyle\sum_{i:ij\in E} x_{ij} = x_j = \sum_{i:ji\in E} x_{ji} \\ \displaystyle\sum_i x_{it} = K = \sum_i x_{si} \end{array}\right\} \mathbf{x} \in \text{FLOW}_K \\ & x_i \,, x_{ij} \quad \text{are} \quad \text{integer.} \end{aligned}$$

The above formulation encodes the joint selection of K tracks using the following selection variables: $x_i \in \{0,1\}$ is a binary indicator variable taking the value 1 when the *detection* i is selected in some track; $x_{ij} \in \{0,1\}$ is a binary indicator variable taking the value 1 when detection i and detection j are *connected* through the *same* track in nearby time frames. The index i ranges over possible detections across the whole video. c_i denotes the cost of selecting detection i in a specific frame (and represents the negative detection confidence) while c_{ij} represents the negative of the correspondence strength between detections i and j. The set of possible connections between detections is represented by E and could be a subset of all pairs of detections in nearby frames by using choice heuristics (such as spatial proximity). The quality of track selection is quantified by the objective in (1).

The constraint $\sum_{i:ij\in E} x_{ij} = x_j = \sum_{i:ji\in E} x_{ji}$, which has the structure of a *flow conservation constraint* [1], encodes the correct claimed semantic that x_{ij} can take the value 1 if and only if both x_i and x_j take the value 1, and moreover, that each detection belongs to *at most one track*, enforcing the fact that two objects cannot occupy the same space. Finally, the constraint $\sum_i x_{it} = K = \sum_i x_{si}$ ensures that exactly K tracks are selected (dummy "source" and "sink" variables with the fixed value $x_s = x_t = K$ are added; the connection variables x_{si} and x_{it} represent the start and end of tracks respectively).

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

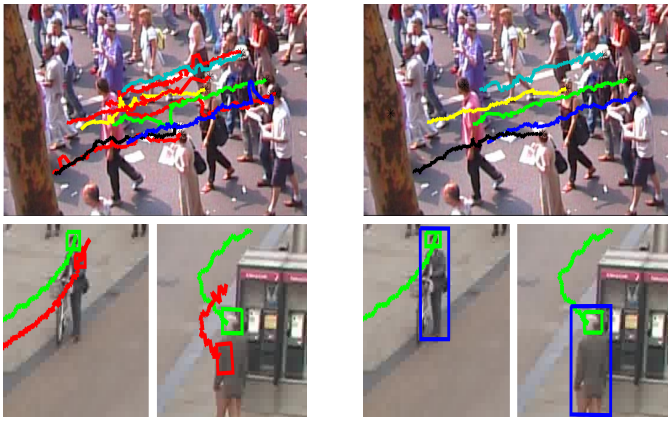

Figure 1: Results of network flow tracking using cost functions with/without pairwise terms. (Top Row): a pairwise term that penalizes the overlap between different tracks helps resolving ambiguous (red tracks) tracks in crowded scenes. (Bottom Row): a pairwise term that encourages the consistency between two signals (here head detections and body detections) helps eliminating failures (red tracks) of object detectors.

To summarize, the above optimization problem can be solved efficiently using existing network flow or linear algebra packages [1] when the integer constraint is relaxed, and provides a convenient framework to transform the tracking problem into a *track selection* problem. We use this conversion as a starting point to add additional constraints and costs on the selection process to influence it in desirable ways to address challenging scenarios.

We make the following contributions in this paper:

- We propose a new *non-greedy* approach to optimize pairwise terms within a min-cost network flow framework. Our solution is generic and allows the simultaneous optimization of any type of pairwise costs.
- We propose a global optimization strategy with a convex relaxation that allows us to minimize pairwise costs using linear optimization, and a principled Frank-Wolfe style rounding procedure to obtain integer solutions with a certificate of suboptimality. The optimization procedure is empirically stable, allowing the practitioner to focus on modeling.
- To illustrate our method, we propose two particular examples of pairwise costs: the first discourages significant overlaps between distinct tracks; the second models the spatial co-occurrence of different types of detections. This allows us to better model complex dynamic scenes with substantial clutter and partial occlusions.
- Using our method, we show improved tracking results on several real-world videos. In addition, we propose a new strategy to evaluate tracking results that better measures the longevity of overlap between output tracks and ground truth.

[1] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice-Hall, Inc., 1993.
[2] Li Zhang, Yuan Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.