Learning Optimal Parameters For Multi-target Tracking

Shaofei Wang
shaofeiw@uci.edu
Charless C. Fowlkes
fowlkes@ics.uci.edu

Dept of Computer Science
University of California
Irvine, CA, USA

Multi-target tracking problems are traditionally tackled in two different ways. One way is to first group detections into candidate tracklets and then perform scoring and association of these tracklets [5, 6], this can be done in either an online/streaming fashion or an offline/batch fashion and it allows tracklets to be scored with richer trajectory and appearance models. Another approach is to attempt to include higher-order constraints directly in a combinatorial framework [1, 2]. In either case, there are a large number of parameters associated with these richer models which become increasingly difficult to set by hand and necessitate the application of machine learning techniques.

In this paper, we describe an end-to-end framework for learning parameters of min-cost flow multi-target tracking problem with quadratic trajectory interactions including suppression of overlapping tracks and contextual cues about co-occurrence of different objects. Our approach utilizes structured prediction with a tracking-specific loss function to learn the complete set of model parameters. Under our learning framework, we evaluate two different approaches to finding an optimal set of tracks under quadratic model objective based on an LP relaxation and a novel greedy extension to dynamic programming that handles pairwise interactions.

In a min-cost flow multi-target tracking problem, the set of optimal (most probable) tracks can be found by solving an integer linear program (ILP) over flow variables $\mathbf{f}$.

$$\min_{\mathbf{f}} \sum_i c_i^s f_i^s + \sum_{ij\in E} c_{ij} f_{ij} + \sum_i c_i f_i + \sum_i c_i^t f_i^t \tag{1}$$

$$\text{s.t.} \quad f_i^s + \sum_j f_{ji} = f_i = f_i^t + \sum_j f_{ij} \tag{2}$$

$$f_i^s, f_i^t, f_i, f_{ij} \in \{0,1\} \tag{3}$$

where E is the set of valid transitions between sites in successive frames. The costs c_i represent the negative log-likelihood ratio of an object appearing at a particular spatio-temporal location i based on image evidence, c_{ij} represents the cost of a transition between a location i in one frame and j in a subsequent frame and c^s and c^t are associated with the birth or death of a track. The flow conservation constraint (2) enforces that a detection at site i can only be active as part of a single contiguous track passing through that location.

It is also possible to capture interactions between multiple tracks by adding a pairwise cost term denoted q_{ij} for jointly activating a pair of flows f_i and f_j corresponding to detections at sites i and j. Adding this term to 1 yields an Integer Quadratic Program (IQP):

$$\min_{\mathbf{f}} \sum_i c_i^s f_i^s + \sum_{ij\in E} c_{ij} f_{ij} + \sum_i c_i f_i + \sum_{ij\in EC} q_{ij} f_i f_j + \sum_i c_i^t f_i^t \tag{4}$$

$$\text{s.t.} \quad (Eq.\,2),\ (Eq.\,3)$$

In our experiments, we investigate pairwise contextual interactions between pairs of sites in the same video frame which we denote by EC. The addition of quadratic terms makes the objective (4) NP-hard. We thus propose a novel greedy approximation based on repeated passes of dynamic programming and compare it with a standard LP relaxation-based approach.

We formulate parameter learning of tracking models as a structured prediction problem. Assume we have N training videos with detector outputs and corresponding ground-truth track associations specified by flow variables $\{(X_n, \mathbf{f}_n)\}$. We can parameterize the network flow costs c as a linear function of image and detection features where the parameters are specified by weight vector $\mathbf{w}$. We propose to discriminatively learn tracking model parameters $\mathbf{w}$ using a structured SVM with margin rescaling:

$$\mathbf{w}^* = \underset{\mathbf{w},\xi_n\geq 0}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_n \tag{5}$$

$$\text{s.t. } \mathbf{w}^T\Psi(X_n,\widehat{\mathbf{f}}) - \mathbf{w}^T\Psi(X_n,\mathbf{f}_n) \geq L(\mathbf{f}_n,\widehat{\mathbf{f}}) - \xi_n \quad \forall n,\widehat{\mathbf{f}}$$

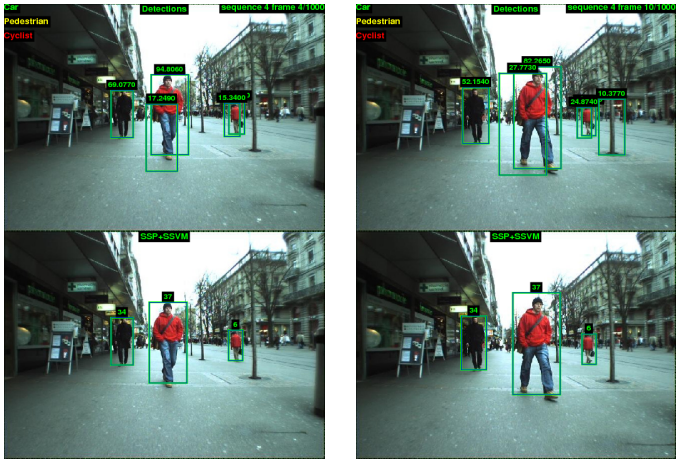

Figure 1: By learning a proper set of parameters, even a basic network-flow model without pairwise potentials can successfully prune away many false tracks by reasoning about detection confidence and transition smoothness.

where $\Psi(X_n,\mathbf{f}_n)$ are the features extracted from nth training video. $L(\mathbf{f}_n,\widehat{\mathbf{f}})$ is a loss function that penalizes any difference between the inferred label $\widehat{\mathbf{f}}$ and the ground truth label $\mathbf{f}_n$ and which satisfies $L(\mathbf{f}_n,\mathbf{f}_n) = 0$.

We use a standard cutting plane approach [4] to optimize the parameters $\mathbf{w}$ by repeatedly performing loss-augmented inference to find flows $\widehat{\mathbf{f}}$ that violate the constraint for each training example. Specifically, we propose to use a decomposable loss L for transition links that attempts to capture important aspects of multi-object tracking accuracy (MOTA) by taking into account the length and localization of transition links rather than using a constant (Hamming) loss on mislabled links.

Implementation details of approximate algorithms as well as the definition of tracking features and loss can be found in our full paper. Surprisingly, we found that with properly learned parameters, even the simple min-cost flow objective (1) yields better results than state-of-the-art methods on challenging MOT and KITTI benchmarks, while the quadratic terms improves the performance even further for tracking with ordinary, multi-category detector such as DPM [3].

[1] Asad A. Butt and Robert T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[2] Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. On pairwise cost for multi-object network flow tracking. *CoRR*, abs/1408.3304, 2014.

[3] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[4] T. Joachims, T. Finley, and Chun-Nam Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[5] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association with online target-specific metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[6] Bo Yang and Ram Nevatia. An online learned crf model for multi-target tracking. In *In CVPR*, 2012.