Target Identity-aware Network Flow for Online Multiple Target Tracking

Afshin Dehghan[1] Yicong Tian[1] Philip. H. S. Torr[2] Mubarak Shah[1]
[1]Center for Research in Computer Vision, University of Central Florida. [2]Department of Engineering Science, University of Oxford.

Online learning methods have been used extensively for tracking deformable objects in the context of *single object tracking* [1]. However, its extension to multiple objects remains relatively unexplored and is limited to only few works [4]. In this paper we propose a tracking method based on online learning, which solve detection and global data association simultaneously. At the core of our framework lies a structured learning which learns a model for each target. The inference is done for all the targets simultaneously in a batch of frames, through a new target identity-aware network flow graph. Despite other online trackers which are temporally local, our method provides the tracks across a segment of a video. The input to our tracker in every frame, is densely sampled candidate windows instead of sparse detections. This allows our tracker to infer temporal consistency between the frames and correct poor detections, thus avoiding error propagation.

Proposed Approach. Given the initial bounding boxes for the objects entering the scene (from annotation or using an object detector), our method starts by training a model for each of the objects through structured learning. During learning, the most violated constraints are found by searching for a set of tracks that minimize the cost function of our target identity-aware network flow. Later, the same network is used to find the best tracks in the next temporal span (segment) of a sequence. The new tracks are later used to update the model through passive aggressive algorithm.

Learning. Given a set of τ training images, $X = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^\tau\} \subset \mathcal{X}$, along with label $Y = \{\mathbf{y}_1^1, \mathbf{y}_2^1, ..., \mathbf{y}_K^1, ..., \mathbf{y}_{K-1}^\tau, \mathbf{y}_K^\tau\} \subset \mathcal{Y}$, where $\mathbf{y}_k^t$, defines the bounding box location of object k in frame t, the target models are obtained through structured learning [3]. The aim of learning is to find a prediction function $f : \mathcal{X} \mapsto \mathcal{Y}$, which directly predicts the locations of all the objects in a set of frames. The task of structured learning is to learn a prediction function of the form

$$f_w(X) = \arg\max_{Y \in \mathcal{Y}} \sum_{t=1}^{\tau} \sum_{k=1}^{K} \mathbf{w}_k^T \phi(\mathbf{x}^t, \mathbf{y}_k^t), \tag{1}$$

where $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K\}$ is the concatenation of the models for all the K objects. $\phi(\mathbf{x}^t, \mathbf{y}_k^t)$ is the joint feature map which represents the feature extracted at location $\mathbf{y}_k^t$ in frame t. Due to exponential possible combination of bounding boxes in $\mathcal{Y}$, exhaustive verification of constraint is not feasible. However [2, 3] showed that high quality solution can be obtained in polynomial time by using only the *most-violated constraints*, i.e a set of bounding boxes that maximize the sum of scores and loss functions. Once the model parameters are learned ($\mathbf{w}$), we use the same inference that we used for finding the *most-violated constraints* to find the best set of tracks for all the K objects in next segment of the video.

Track Inference. Given the model parameters, $\mathbf{w}$, and *dense overlapping bounding boxes* in each frame, the goal is to find a set of candidate windows, called a track, for each object which maximizes the score in Eq. 1. We propose to formulate the inference as a global data association which helps reducing the search space by enforcing some temporal consistency across the candidates in consecutive frames. We propose a new network called Identity-Aware network, which is shown in Fig. 1. The black circles represent all possible candidate locations in each frame (densely sampled across the entire frame). Each candidate location is represented with a pair of nodes that are linked through K *observation edges*; one *observation edge* for each identity. Our network has K sources and K sinks, each belonging to one object. The rest of the network is similar to that of traditional network flow.

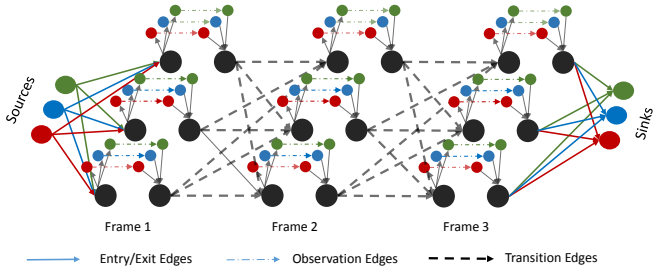

Figure 1: Shows the network used in our inference for three identities. Each identity is shown with a unique color. The flow entering each node can take only one of the three observation edges depending on which source (identity) does it belong to. The constraint in Eq. 5 ensures that one candidate can belong to only one track, so the tracks will not overlap.

Target Identity-aware Network Flow. Once the graph $G(V,E)$ is constructed, our aim is to find a set of K flows (tracks) by pushing a unit of flow through each source node. The flow $f_{i,j}^k$, is found by minimizing the following cost function:

$$C(f) = \sum_{k=1}^{K} \sum_{(i,j) \in E} c_{ij}^k f_{ij}^k. \tag{2}$$

The flow passing through these edges need to satisfy some constraints to ensure that it can actually represent a track in a real world. The set of constraints that we define in our graph are as follow:

$$\sum_j f_{ij}^k - \sum_j f_{ji}^k = \begin{cases} 1 & \text{if } i = s_k \\ -1 & \text{if } i = t_k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$f_{ij}^k \geq 0 \qquad \forall (i,j) \in E \quad and \quad 1 \leq k \leq K \tag{4}$$

$$\sum_{k=1}^{K} f_{ij}^k \leq 1 \tag{5}$$

The above optimization problem can be formulated as an Integer Program problem. However, our experiments show that without pruning steps, which reduces the number of candidate windows, it is intractable to find a solution for a large number of people in a long temporal span. In this paper, we propose a Lagrange relaxation solution to this problem. We show that after relaxing the hard constraints, the problem in each iteration, reduces to finding the best track for each target separately. The global solution to this can be found in linear time through dynamic programming. Moreover, our iterative optimization allows us to incorporate spatial constraint which further improves the tracking results. The detail explanation of our Lagrange relaxation optimization as well as spatial constraint are described in the paper. Our work is one of the very few attempts that aims to solve tracking multiple objects by solving detection and tracking simultaneously. We hope that our results encourage other researcher to discover this direction more

[1] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011.

[2] Chun nam Yu and Thorsten Joachims. Learning structural svms with latent variables. In *ICML*, 2009.

[3] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *JMLR*, 2005.

[4] L. Zhang and L.J.P. van der Maaten. Structure Preserving Object Tracking. In *CVPR*, 2013.

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