Joint Tracking and Segmentation of Multiple Targets

Anton Milan[1], Laura Leal-Taixé[2], Konrad Schindler[2], Ian Reid[1]
[1]University of Adelaide, Australia. [2]Photogrammetry and Remote Sensing Group, ETH Zürich.

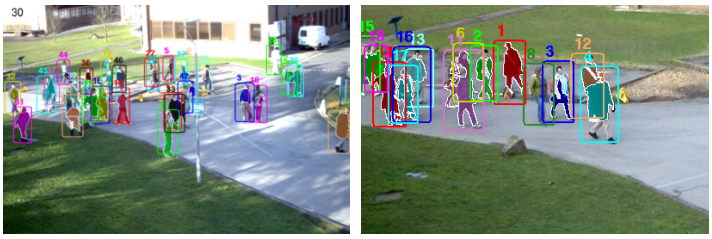

Figure 1: Qualitative tracking and segmentation results.

Tracking-by-detection traditionally relies on a set of sparse detections that serve as input to a high-level tracker whose goal is to correctly associate these "dots" over time. An obvious shortcoming of this approach is that most information available in image sequences is simply ignored by thresholding weak detection responses and applying non-maximum suppression. We argue that it is beneficial to consider *all* image evidence to handle tracking in crowded scenarios. In contrast to many previous approaches, we aim to assign a unique target ID not only to each individual detection, but to every (super-)pixel in the entire video (*cf.* Fig. 1). This low-level information enables us to recover trajectories of largely occluded targets since the partial image evidence of the superpixels often persists even in the absence of detections. Exploiting low-level information in the context of multi-target tracking has been recently proposed [2, 3]. However, one major limitation of previous approaches is their inherent inability to track targets through full occlusions. In addition, a target's state (*i.e.* its location) is only defined implicitly by the segmentation [2], which makes it rather difficult to estimate the full extent in case of (partial) occlusion. Our experiments confirm that these methods show relatively poor performance in crowded scenes when evaluated with standard multi-target tracking measures. This work overcomes both limitations by explicitly modelling the continuous state of all targets throughout the entire sequence.

In common with some other approaches [5–7] we formulate the problem as one of finding a set of continuous trajectory hypotheses that best explains the data, but our approach differs in that we take account of the low-level information in scoring the trajectory hypotheses. We do this by modelling the problem as a multi-label conditional random field (CRF). Furthermore, contrary to prior closely related work [1, 6], our trajectory model is not a simple space-time curve but rather a volumetric tube with a rectangular cross-section, allowing for a more accurate representation. Our method shows encouraging results on many standard benchmark sequences and significantly outperforms state-of-the-art tracking-by-detection approaches in crowded scenes with long-term partial occlusions.

Our high-level approach to this problem follows a model selection strategy, similar to [1]: we generate an overcomplete set of trajectory hypotheses and then optimize an objective that chooses which hypotheses participate in the solution. This objective must capture agreement with image evidence along with our prior beliefs about the properties of valid trajectories such as their continuity, dynamics, *etc.* We formulate the assignment of detections and (super)-pixels to trajectory hypotheses as a multi-label conditional random field (CRF) with nodes $\mathcal{V} = \mathcal{V}_S \cup \mathcal{V}_D$ and edges $\mathcal{E}$, where $\mathcal{V}_S$ represents all superpixel nodes and $\mathcal{V}_D$ all detection nodes (see Fig. 2). Each random variable $v \in \mathcal{V}$ can take on a label from the label set $\mathcal{L} = \{1, \ldots, N, \varnothing\}$, which can be either a unique target ID or the background (false alarm) label $\varnothing$. We aim to find the most probable labelling $\mathbf{v}^*$ for all nodes given the observations, which is equivalent to minimizing the corresponding Gibbs energy: $\mathbf{v}^* = \arg\min_{\mathbf{v}} E(\mathcal{V})$. We define the energy as follows:

$$E(\mathcal{V}) = \sum_{d \in \mathcal{V}_D} \phi^{\mathcal{V}_D}(d) + \sum_{s \in \mathcal{V}_S} \phi^{\mathcal{V}_S}(s) + \sum_{(v,w) \in \mathcal{E}} \psi(v,w) + \psi^{\lambda}, \qquad (1)$$

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

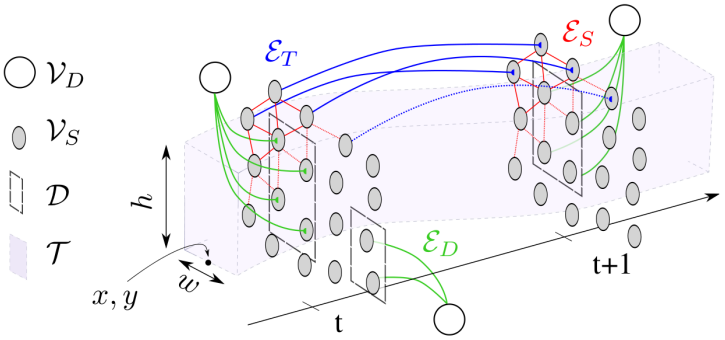

Figure 2: Our CRF model for two consecutive frames, showing superpixel nodes $\mathcal{V}_S$, detection nodes $\in \mathcal{V}_D$ and only a subset of the pairwise edges $\mathcal{E}$.

with unaries $\phi^{\mathcal{V}_D}$ and $\phi^{\mathcal{V}_S}$, pairwise potentials ψ, and a regulariser ψ^{λ}. Detection unaries model the overlap between a detection bounding box and a trajectory hypothesis $\mathcal{T}$. Superpixel unaries capture the colour consistency between a superpixel and a foreground model, as well as the motion consistency *w.r.t.* a trajectory hypothesis.

The MDL term ψ^{λ} restricts the number of trajectories from growing arbitrarily high. Because the number of targets is typically unknown, it is necessary to include this regulariser that favours solutions with fewer labels. In our formulation, this global factor also acts as a trajectory-specific prior, capturing target dynamics, the hypothesis shape and size, track persistence, and the foreground likelihood covered by a hypothesis. By involving the pixel information in the optimization we enable the label IDs to persist even when there is no explicit detector evidence. Tab. 1 shows results of our approach compared to two top public submissions on the recent MOTChallenge benchmark. Further model details and more experiments can be found in the paper.

Our conclusion is that exploiting all image information helps to improve multiple target tracking in regions of long-term partial occlusions. Moreover, our joint tracking and segmentation framework provides reasonable instance-based segmentation masks in crowded scenarios.

Table 1: Results on the MOTChallenge 2015 Benchmark.

Method	TA	TP	Rcll	Prcn	MT	ML	ID	FM
RMOT [8]	18.6	69.6	40.0	66.4	5.3	53.3	**684**	1282
CEM [6]	19.3	70.7	**43.7**	65.4	**8.5**	**46.5**	813	1023
MotiCon [4]	**23.1**	70.9	41.7	71.1	4.7	52.0	1018	1061
SegTrack	22.5	**71.7**	36.5	**74.0**	5.8	63.9	697	**737**

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR 2012*.

[2] S. Chen, A. Fern, and S. Todorovic. Multi-object tracking via constrained sequential labeling. In *CVPR 2014*.

[3] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi. Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In *ECCV 2012*, volume 7576, pages 552–565.

[4] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *CVPR 2014*.

[5] B. Leibe, K. Schindler, and Luc Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV 2007*.

[6] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multi-target tracking. *IEEE T. Pattern Anal. Mach. Intell.*, 36(1):58–72, 2014.

[7] D. Mitzel, E. Horbert, A. Ess, and B. Leibe. Multi-person tracking with sparse detection and continuous segmentation. In *ECCV 2010*, pages 397–410.

[8] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *WACV*, 2015.