Online Domain Adaptation for Multi-Object Tracking

Adrien Gaidon
adrien.gaidon@xrce.xerox.com
Eleonora Vig
eleonora.vig@xrce.xerox.com

Computer Vision Group
Xerox Research Centre Europe
Meylan, France

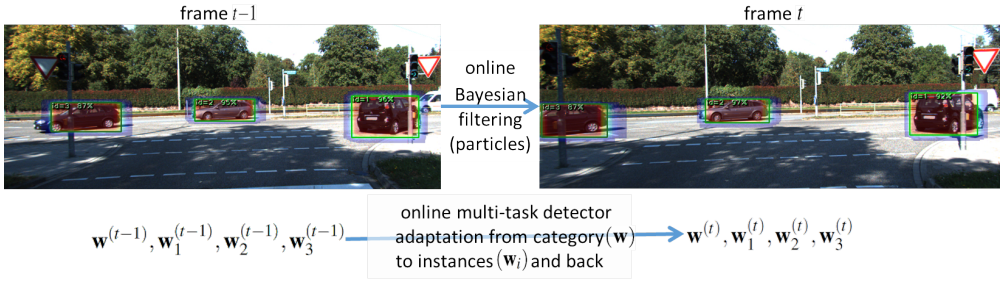

Figure 1: Online domain adaptation for MOT via Bayesian filtering coupled with multi-task adaptation of all detectors jointly.

Automatically detecting, labeling, and tracking objects in videos depends above all on accurate category-level object detectors. These might, however, not always be available in practice, as acquiring high-quality large scale labeled training datasets is either too costly or impractical for all possible real-world applications. A scalable solution consists in re-using object detectors pre-trained on generic datasets. This work is the first to investigate the problem of *on-line domain adaptation of object detectors for causal multi-object tracking (MOT)*. We propose to alleviate the dataset bias by adapting detectors from *category to instances, and back*: (i) we jointly learn all target models by adapting them from the pre-trained one, and (ii) we also adapt the pre-trained model on-line. Previous works investigated detector adaptation *or* on-line learning of appearance models, but not both *jointly*. Our approach can be interpreted as a generalization. We integrate our domain adaptation strategy in a novel motion model combining learned deterministic models with standard Bayesian filtering (*cf.* figure above) inspired by the popular Bootstrap filter. In particular, we leverage several techniques not widely used in MOT yet: (i) recent improvements in object detection based on object proposals, (ii) large-displacement optical flow estimation, (iii) the Fisher Vector representation, and (iv) ConvNet features for object detection. In addition, we use a Sequential Monte Carlo algorithm to approximate the filtering distribution of our Markovian motion model of the latent target locations.

Contrary to common practice in MOT, we here use object proposals, which we rank with a category-specific linear classifier parameterized by a vector $\mathbf{w}$. This classifier returns the probability that a candidate window $\mathbf{x}$, represented by a feature vector $\phi_t(\mathbf{x})$, contains an object of the category of interest at time t by $P(\mathbf{x}|\mathbf{w}) = \left(1+e^{-\mathbf{w}^T\phi_t(\mathbf{x})}\right)^{-1}$. To represent proposals, we explore two common representations adapted to the computational constraints of tracking: Fisher Vectors with a single Gaussian and features from the memory-efficient pre-trained GoogLeNet ConvNet [6].

We propose a *convex multi-task learning objective* to *jointly adapt on-line* (i) all trackers from the pre-trained generic detector (*category-to-instance*), and (ii) the pre-trained category-level model from the trackers (*instances-to-category*). The first category-to-instance adaptation happens at the creation of a new track $\mathbf{w}_i^{(t_0)}$ by *warm-starting* its optimization from the category-level model $\mathbf{w}^{(t_0)}$, *i.e.* an already good solution. This leads to faster convergence and stronger regularization. The second category-to-instance adaptation consists in *updating all target models jointly* using multi-task learning. Given the stacked target models $\mathbf{W}^{(t)} = \{\mathbf{w}_1^{(t)}, \ldots, \mathbf{w}_{N_t}^{(t)}\}$ (N_t in total), and the training samples and labels $(\mathbf{X}^{(t)}, \mathbf{y}^{(t)})$ mined for all targets in frame t, updating all appearance models jointly amounts to minimizing the regularized empirical risk:

$$\mathbf{W}^{(t)} = \arg\min_{\mathbf{W}} L_t(\mathbf{X}^{(t)}, \mathbf{y}^{(t)}, \mathbf{W}) + \lambda\,\Omega_t(\mathbf{W}) \qquad (1)$$

with the loss L_t and multi-task regularization term Ω_t defined as:

$$L_t(\mathbf{X}^{(t)}, \mathbf{y}^{(t)}, \mathbf{W}) = \frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{n_i}\sum_{k=1}^{n_i}\ell_t(\mathbf{x}_{i,k}\ ,\ y_{i,k}\ ,\ \mathbf{w}_i) \qquad (2)$$

method	MOTA↑	MOTP↑	MT↑	ML↓	Rec.↑	Prec.↑	FAR↓	IDS↓	FRG↓
DP_MCF† [5]	12.0%	68.5%	0.1%	80.2%	14.6%	**85.5%**	**7.7%**	**84**	**327**
G_TBD† [2]	17.5%	68.0%	0.9%	59.2%	30.0%	71.3%	37.6%	115	528
CFT [4]	17.6%	66.7%	1.8%	45.7%	33.5%	69.1%	47.2%	238	592
CIT [3]	22.8%	68.5%	**1.9%**	**43.4%**	33.9%	76.5%	32.6%	380	809
ODAMOT	**23.6%**	**68.7%**	1.8%	43.6%	**34.2%**	77.5%	31.1%	376	784

Table 1: MOT results on the PASCAL-to-KITTI domain adaptation dataset for the R-CNN-like detector. Methods with † are offline, the others are online.

$$\Omega_t(\mathbf{W}) = \frac{1}{2N_t}\sum_{i=1}^{N_t}\|\mathbf{w}_i - \bar{\mathbf{w}}^{(t-1)}\|_2^2, \qquad (3)$$

where $\bar{\mathbf{w}}^{(t-1)}$ is the (running) mean of all previous instance models, and $\ell_t(\mathbf{x}, y, \mathbf{w})$ is the logistic loss.

The *instance-to-category* adaptation allows to continuously specialize the global appearance model to the specific video stream. Once the detectors $\mathbf{w}_i$ are updated in frame t, a new scene-adapted category detector is readily available as the running average of instance models:

$$\bar{\mathbf{w}}^{(t)} = \frac{1}{\bar{N}_{t-1}+N_t}\left(\bar{N}_{t-1}\bar{\mathbf{w}}^{(t-1)} + \sum_{i=1}^{N_t}\mathbf{w}_i^{(t)}\right), \text{ where } \bar{N}_{t-1} = \sum_{j=1}^{t-1}N_j. \qquad (4)$$

Our multi-task formulation enforces parameter sharing between models to reduce model drift and robustly handle false alarms, while allowing for continuous domain adaptation to gradually decrease missed detections.

We evaluate our algorithm (ODAMOT) on the challenging KITTI car tracking benchmark [1]. On this dataset, ODAMOT achieves 57.06% MOTA and ranks third of all published methods despite its simple occlusion reasoning. We then quantitatively measure the benefit of our domain adaptation strategy on the new PASCAL-to-KITTI dataset we introduce to study the domain mismatch problem in MOT. The training set (the *source domain*) of this dataset consists of the training images of the standard Pascal VOC 2007 detection challenge, whereas the test set (the *target domain*) includes the 21 training videos of the KITTI tracking challenge. As expected, unrelated training data degrades MOT performance (*cf.* Table 1), however, our results show that domain adaptation partly mitigates this problem. Our multi-task adaptation from category-to-instances and back allows to improve overall MOT accuracy by increasing recall while maintaining high precision and limiting model drift.

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[2] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D Traffic Scene Understanding from Movable Platforms. *PAMI*, 2014.

[3] D. Hall and P. Perona. Online, Real-Time Tracking Using a Category-to-Individual Detector. In *ECCV*, 2014.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-Learning-Detection. *PAMI*, 2011.

[5] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.