Self-Learning Camera: Autonomous Adaptation of Object Detectors to Unlabeled Video Streams

Adrien Gaidon, Eleonora Vig, Gloria Zen, Jose A. Rodriguez-Serrano

Xerox Research Centre Europe

Motivation

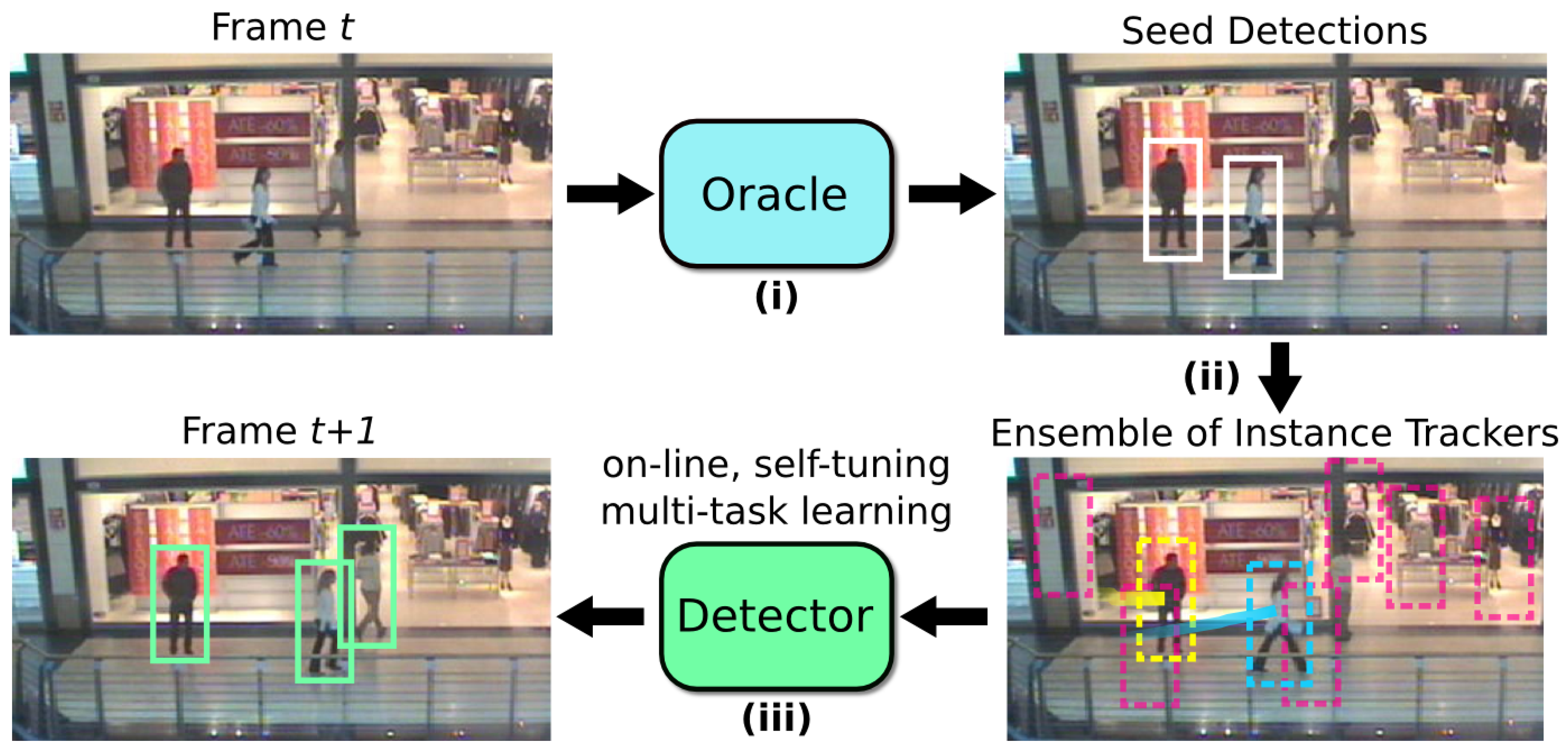
- Learning object detectors requires massive amounts of labeled training data from the source of interest
- $\rightarrow$ Impractical with:
  - many different sources (network of cameras)
  - constantly changing sources (mobile cameras)

Goal – Autonomous self-learning of detectors

Online unsupervised learning of detectors that

- continuously adapt to streaming data sources
- without any labeled data
- without manually set hyperparameters

Overview – Ensemble of Instance Trackers



i Generate seed detections from a confident but laconic oracle

ii Jointly learn instance-level models using online multi-task learning: **Ensemble of Instance Trackers (EIT)**

iii Generate a category-level model from instance models

iv Mine for new training examples

Learning a detector

- An object detector (parameterized by $\mathbf{w}$) assigns an image window $\mathbf{x}$ represented by a feature vector $\phi(\mathbf{x})$ to a category with the probability:

$$P(\mathbf{x}) = \left(1 + e^{-\left(\mathbf{w}^T\phi(\mathbf{x})+b\right)}\right)^{-1} \quad (1)$$

- Training data: tracking-by-detection of seeds $\rightarrow$ pool of candidate locations: one positive (closest match) + hard negative examples

Multi-task learning of an ensemble of instance trackers

- **Ensemble of Instance Trackers (EIT)**: set of N object-instance detectors in the current frame parameterized by $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$
- Jointly minimize over the ensemble parameters:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} L(\mathbf{X}, \mathbf{y}, \mathbf{W}) + \lambda\Omega(\mathbf{W}), \quad (2)$$

where $L(\mathbf{X}, \mathbf{y}, \mathbf{W}) = \sum_{i=1}^{N}\sum_{k=1}^{n_i} \ell(\mathbf{x}_k^{(i)}, y_k^{(i)}, \mathbf{w}_i)$ is logistic loss and $(\mathbf{x}_k^{(i)}, y_k^{(i)})$ are training samples of object i (n_i in total)

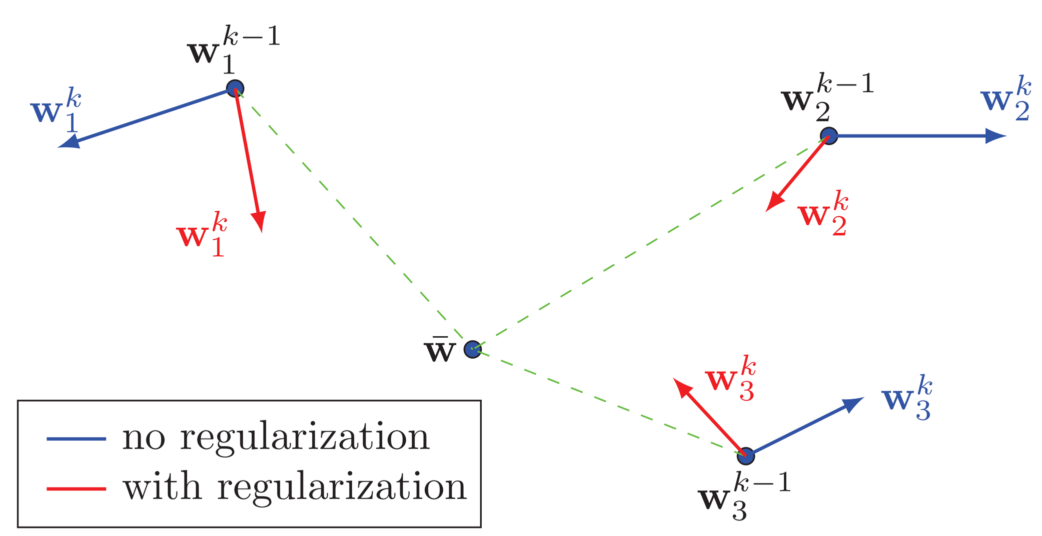
- Use a multi-task regularization term:

$$\Omega(\mathbf{W}) = \frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{w}_i - \bar{\mathbf{w}}\|_2^2, \quad (3)$$

where $\bar{\mathbf{w}}$ is the (running) mean of all instance models

- Regularization promotes similarity between instance models $\rightarrow$ avoids overfitting and drifting
- New scene-adapted category-level detector $\bar{\mathbf{w}}^*$ = updated mean

Continuous self-tuning online adaptation



- Averaged Stochastic Gradient Descent (ASGD) is used to solve Eq. (2)
- Update rule for each model $\mathbf{w}_i$:

$$\mathbf{w}_i^k = \mathbf{w}_i^{k-1} - \eta\left(\frac{\partial\ell}{\partial\mathbf{w}}(\mathbf{x}_k^{(i)}, y_k^{(i)}, \mathbf{w}_i^{k-1}) + \frac{\lambda}{N}\left(\mathbf{w}_i^{k-1} - \bar{\mathbf{w}}\right)\right), \quad (4)$$

with η learning rate and training samples $(\mathbf{x}_k^{(i)}, y_k^{(i)})$, $k = 1, \ldots, n_i$

- Self-tuning the parameters: greedy search for least-overfitting parameter values that optimize the rank of the closest detection in the current frame

References

- A. Gaidon, G. Zen, J. A. Rodriguez-Serrano, "Self-Learning Camera: Autonomous Adaptation of Object Detectors to Unlabeled Video Streams", *arXiv*, 2014.
- P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, 2010.
- X. Wang, G. Hua, and T. Han, "Detection by detections: Non-parametric detector adaptation for a video.," *CVPR*, 2012.

Results

Video object detection datasets:

	frame size	fps	#frames	class	#objects
CAVIAR (OIs1)	576 × 768	25	295	pedestrian	438
CAVIAR (OIs2)	576 × 768	25	1119	pedestrian	290
CAVIAR (Osow1)	576 × 768	25	1377	pedestrian	2402
CAVIAR (OIsr2)	576 × 768	25	560	pedestrian	811
CAVIAR (Ose2)	576 × 768	25	2725	pedestrian	1737
VIRAT-0401	1080 × 1920	30	58K	car	375K

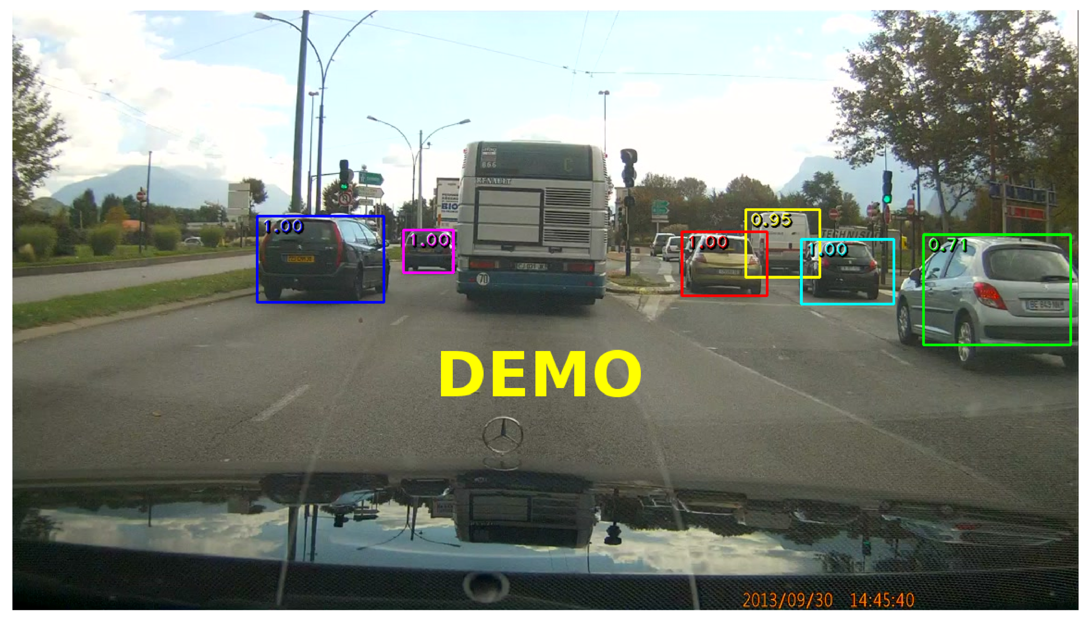
Blackbox oracle: off-the-shelf DPM [2] pretrained on Pascal VOC 2007

Learned detector: Fisher Vectors + approximate sliding window

Quantitative results (Average Precision):

	OIs1	OIs2	OIsr2	Osow1	Ose2	VIRAT-0401
DPM [2]	30.4	52.4	34.9	52.2	34.8	47.0
DbD [3]	**32.1**	56.3	43.1	47.0	**40.9**	N.A.
I-EIT	27.4	53.6	40.6	51.9	38.9	53.1
EIT	29.3	**58.0**	**43.7**	**53.1**	38.1	**53.7**

Qualitative results **(demo)**: use the learned detector for **continuously adapted tracking** of vehicles in videos acquired by mobile cameras



Conclusions

- Continuous category-level learning of object detectors along a data stream
- Mining of positives and (hard) negatives using spatio-temporal structure
- Online multi-task learning of a category model from instances
- Autonomous adaptation over time through self-tuning of hyperparameters

adrien.gaidon@xrce.xerox.com

eleonora.vig@xrce.xerox.com