Learning a Robust Society of Tracking Parts using Co-occurrence Constraints

Elena Burceanu
elena.burceanu@gmail.com
PhD - 2nd year, 1st semester report

April 13, 2018

Tracking

- The root for (m)any video applications (robotics, medical-posture apps, self-driving cars, smart houses, surveillance, describe videos in natural language)
- Generic class tracking:
  - the only GT is the bounding box of the object in the first frame
  - not knowing in advance the properties of the object being tracker (appearance model or motion patterns)
- Challenges in tracking
  - integrate changes in appearance, but keep the model learned so far
  - problems: background clutter, fast or complex motion, deformation, etc
  - drifting: accumulating small errors (eg. bkg as positive sample)
  - decide bounding box based on detection map (weight and height)

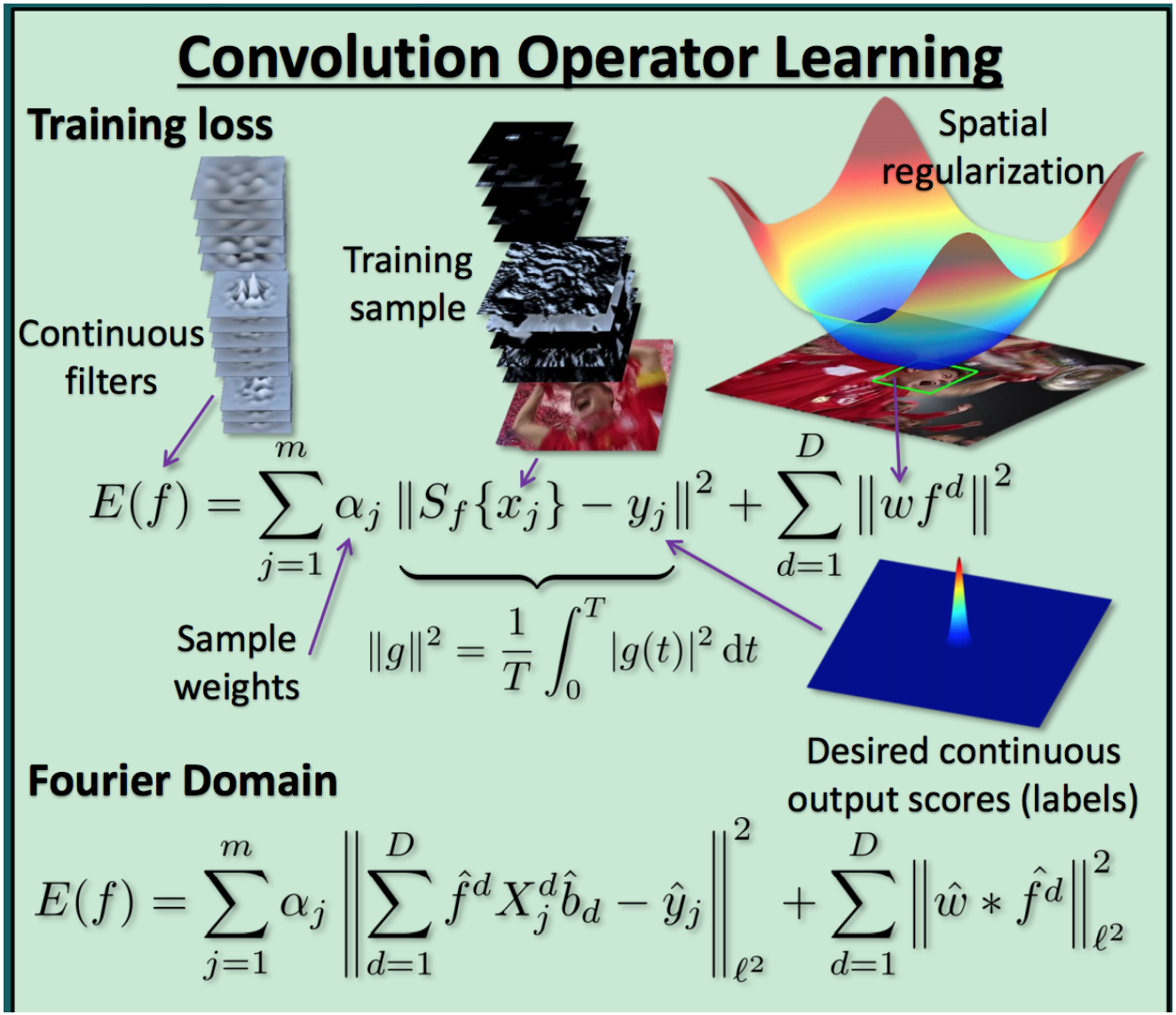
Previous Approaches

- Discriminative Tracking
  - training samples labeled by the tracker
  - rely on the quality of the training set
  - very similar positive samples
  - bad samples (due to occlusion, clutter)
  - results in drifting
- DCF - Discriminative Correlation Filter Trackers
  - efficiently exploits all cyclic shifts of the training samples
  - used at multi-channel level
  - single-resolution feature map
  - learns a set of discrete filters for target localization
  - outputs discrete detection scores

SOTA Trackers

- CCOT (Danelljan et al., 2016)
  - features from multiple layers of CNN
  - **continuous operator** transforms features from discrete to continuous space
  - multi-resolution feature maps would generate artifacts (if combining discrete values)
  - **convolute operator** convolves continuous features with continuous learned filters
  - optimizes computation (Fourier domain)
  - train on each frame (Conjugate Gradient)
- ECO (Danelljan et al., 2017)
  - starts from CCOT
  - reduces the overfit by grouping similar frames (learns a Gaussian mixture model for frames)
  - speedup (factorizes the convolution operation, having fewer filters)

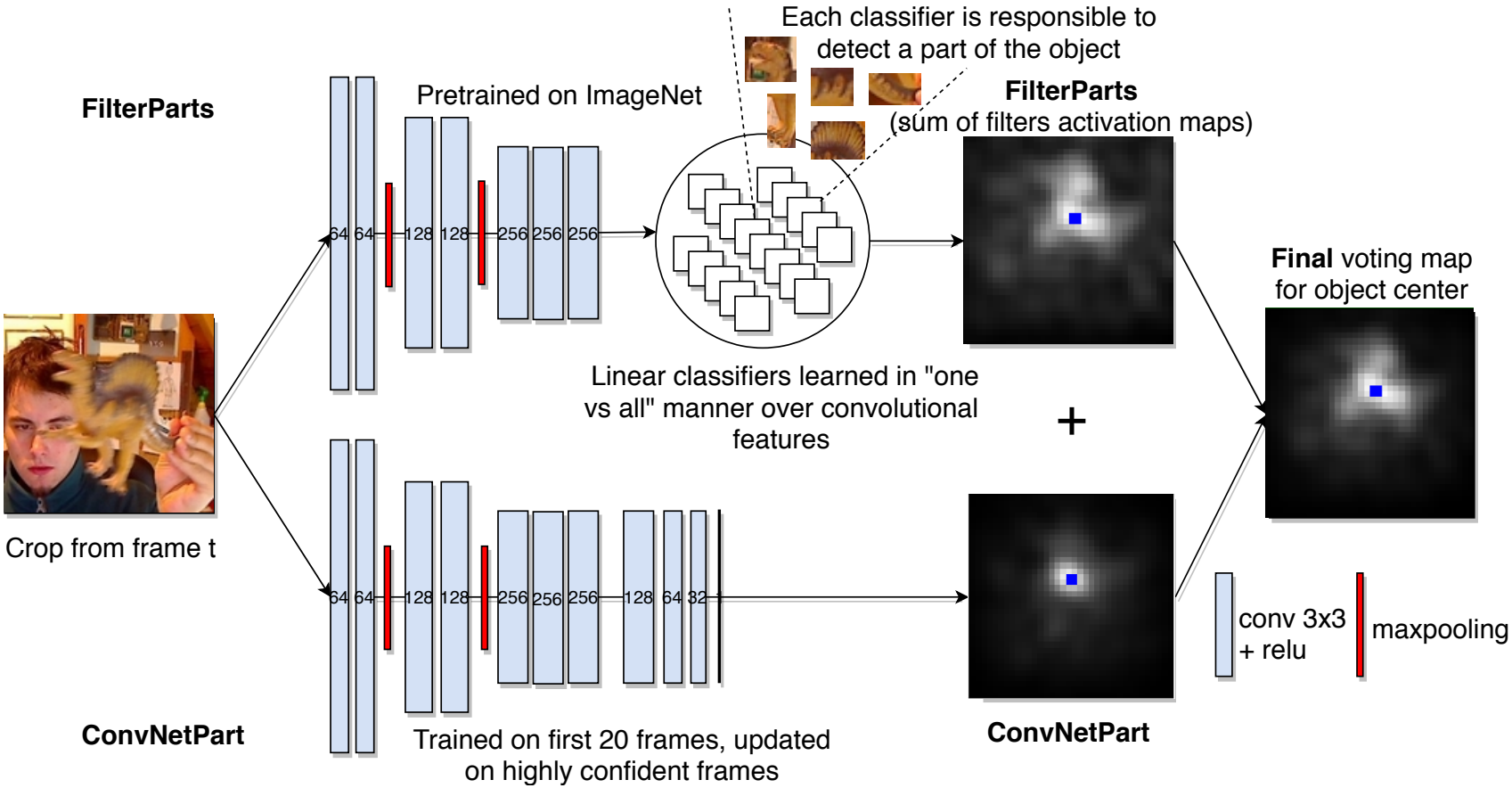
SOTA Trackers



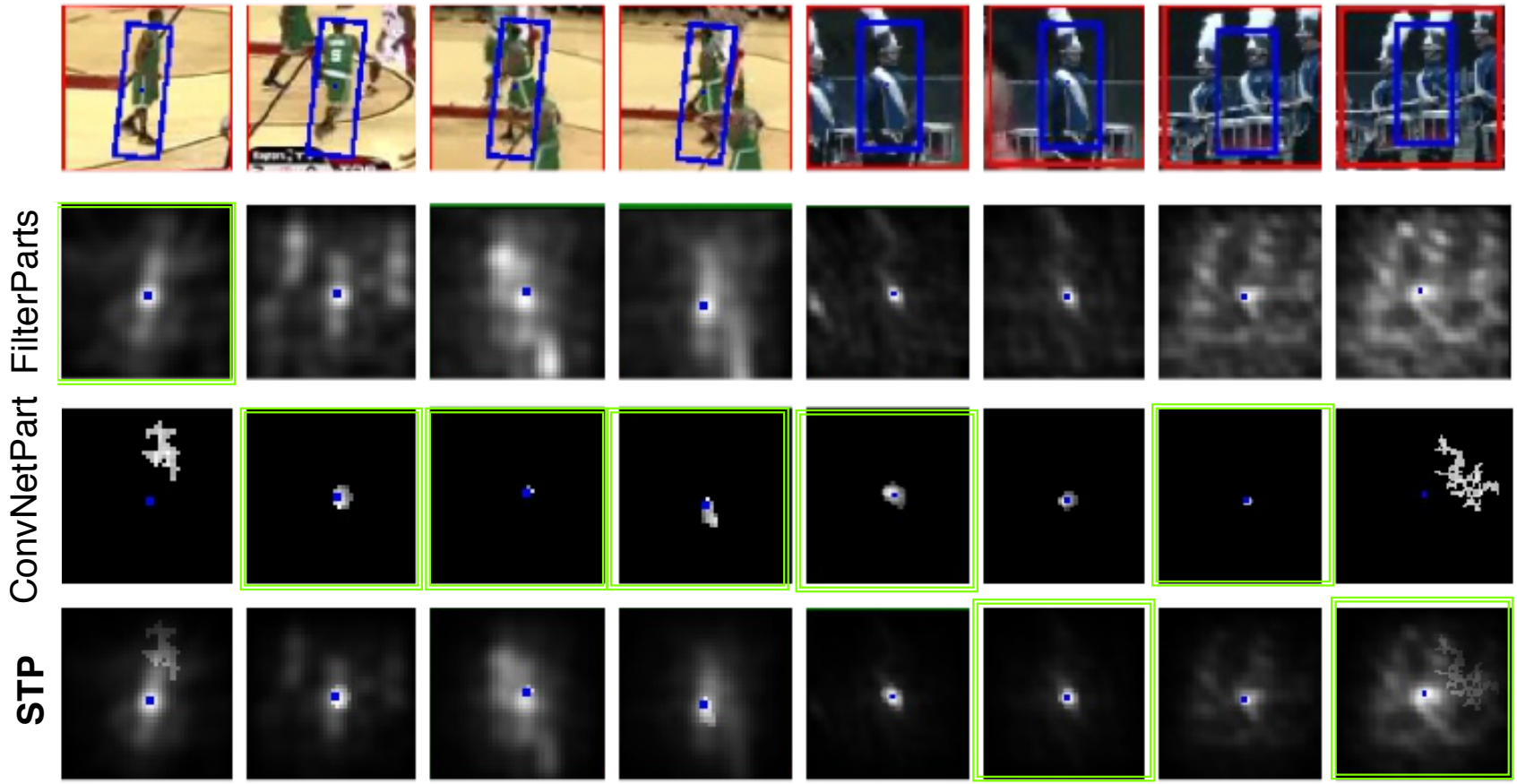
Intuition and Motivation

- Purpose and Approach:
  - adapt the current knowledge to continuous changes
  - robust against background noise
  - many parts, with degrees of complexity, guided by the co-occurrences of their responses
- **1) Stability through steadiness**
  - candidate-reliable-gold (vote co-occurred frequently enough at the same location with the majority)
  - keep the first (very confident) frames
  - uncertainty mask over previous center location
- **2) Robust Adaptation**
  - continuously adapt by validate parts using a temporal buffer
  - update the ConvNetPart on HCF (accumulated over time)
- **3) Robust frame to frame tracking**
  - final vote: peak of the voting map: sum over all parts
  - strong **co-occurrences** of votes at a single location

Society of Tracking Parts



Voting maps



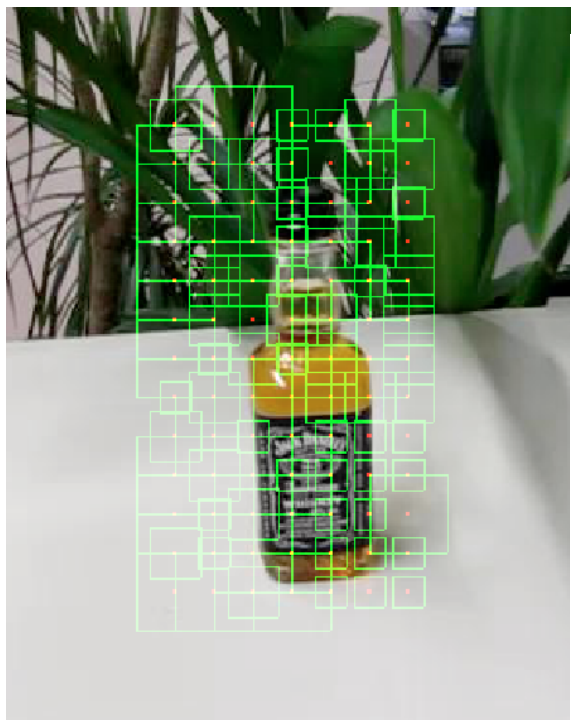
- FilterParts vote is stable
- ConvNetPart has usually a better quality, but sometimes it fails
- the combination provides a more robust maximum

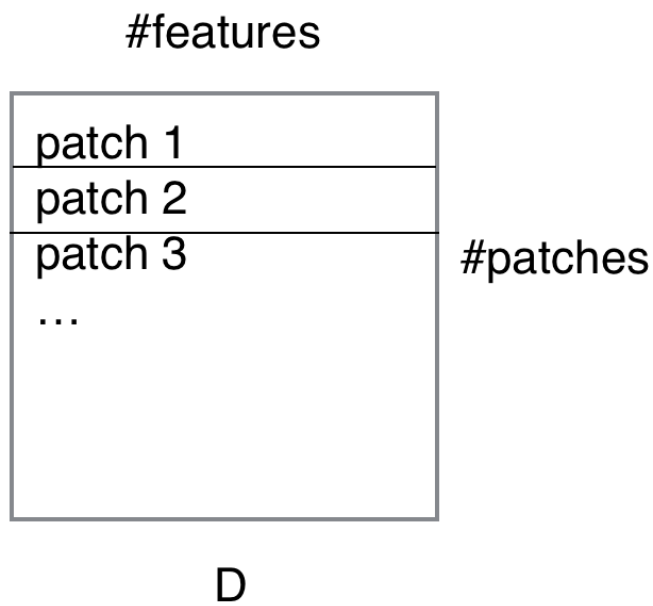
Learning FilterParts I

- Reliability states
  - reliability = frequency at which the maximum activation of a given part is in the neighborhood of the maximum in the final activation
  - every U frames, measure the reliability of a given part
  - promote parts with a large reliability
  - from candidate state (C) to reliable state (R) and to gold (G)
  - remove parts that do not pass the test (except for permanent - Gold - ones)
- FilterParts update phase
  - new parts as candidates
  - classifiers, of different sizes and locations
  - linear filters over activation maps of deep features

Learning FilterParts II

- Choose patches
  - centered on a thin grid over the searching zone
  - build data matrix D (with one patch per row)





- Build linear "1 vs all" classifiers
  - $c_i = (\mathbf{D}^\top \mathbf{D} + \lambda \mathbf{I}_k)^{-1} \mathbf{D}^\top \mathbf{y}_i$

Learning FilterParts III

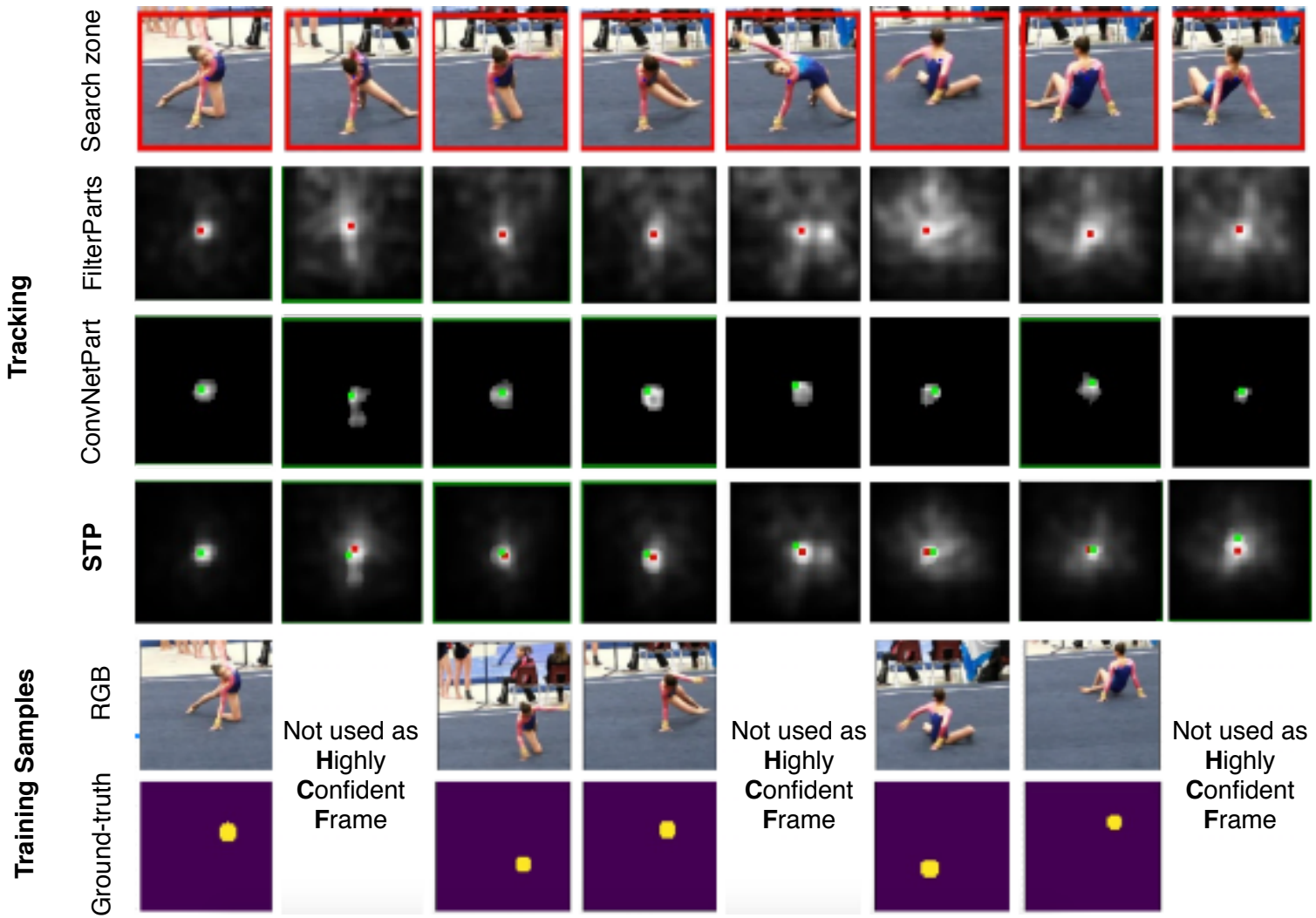
- balance positive vs negatives: weighted linear ridge regression, in closed form: $\theta_i = (\mathbf{D}^\top \mathbf{W}_i \mathbf{D} + \lambda \mathbf{I}_k)^{-1} \mathbf{D}^\top \mathbf{W}_i \mathbf{y}_i$
- $y_i^\top = \begin{bmatrix} 0 & 0 & \dots & 1 & \dots & 0 & 0 \end{bmatrix}$
- **novelty**: for "1 vs all" case, the solution vector (θ_i) has the same direction with the one for the linear ridge regression (c_i), having the ratio $q_i = \frac{n}{1+(n-1)\mathbf{d}_i^\top \mathbf{c}_i}$, so $\theta_i = q_i c_i$

- Advantages
  - c_i can be computed in one operation for all "i"s (not possible for the weighted case)
  - bonus: invert a smaller matrix ($\mathbf{D}\mathbf{D}^\top$ instead of $\mathbf{D}^\top\mathbf{D}$)
  - $c_i = \mathbf{D}^\top(\mathbf{D}\mathbf{D}^\top + \lambda \mathbf{I}_n)^{-1}\mathbf{y}_i$ [1], invert a matrix with 2 orders of magnitude smaller

- we can compute all positive and all negative classifiers with **only one** (**small**) matrix inversion

[1] Matrix Inversion Lemma, see Murphy (2012), Ch. 4.3.4.2

Learning the ConvNetPart



VOT Benchmark

- VOT16, VOT17 benchmarks
  - Single-object, single-camera, model-free, short-term, causal tracker
  - Bounding box: Rotated rectangle
  - Evaluation:
    - Reinitialize tracker after each complete fail
    - Robustness: number of times a tracker drifts of the target
    - Accuracy: average overlap during successful tracking
    - Metric for both: Expected Average Overlap (EAO)
  - 70 trackers in the VOT16 benchmark

Tracker \ Dataset	VOT17			VOT16		
	EAO	Fail rate	Acc	EAO	Fail rate	Acc
STP (ours)	**0.309**	**0.765**	0.44	0.361	**0.47**	0.48
CFWCR	0.303	1.2	0.48	**0.39**	0.81	**0.58**
ECO	0.28	1.13	0.48	0.374	0.72	0.54
CCOT	0.267	1.31	0.49	0.331	0.85	0.52
Staple	0.169	2.5	**0.53**	0.295	1.35	0.54
ASMS	0.169	2.23	0.494	0.212	1.925	0.503
CCCT	-	-	-	0.223	1.83	0.442
EBT	-	-	-	0.291	0.9	0.44
CSRDCF	0.256	1.368	0.491	-	-	-
MCPF	0.248	1.548	0.510	-	-	-
ANT	0.168	2.16	0.464	-	-	-

- best failure rate by large margins (42%-67%), low overlap

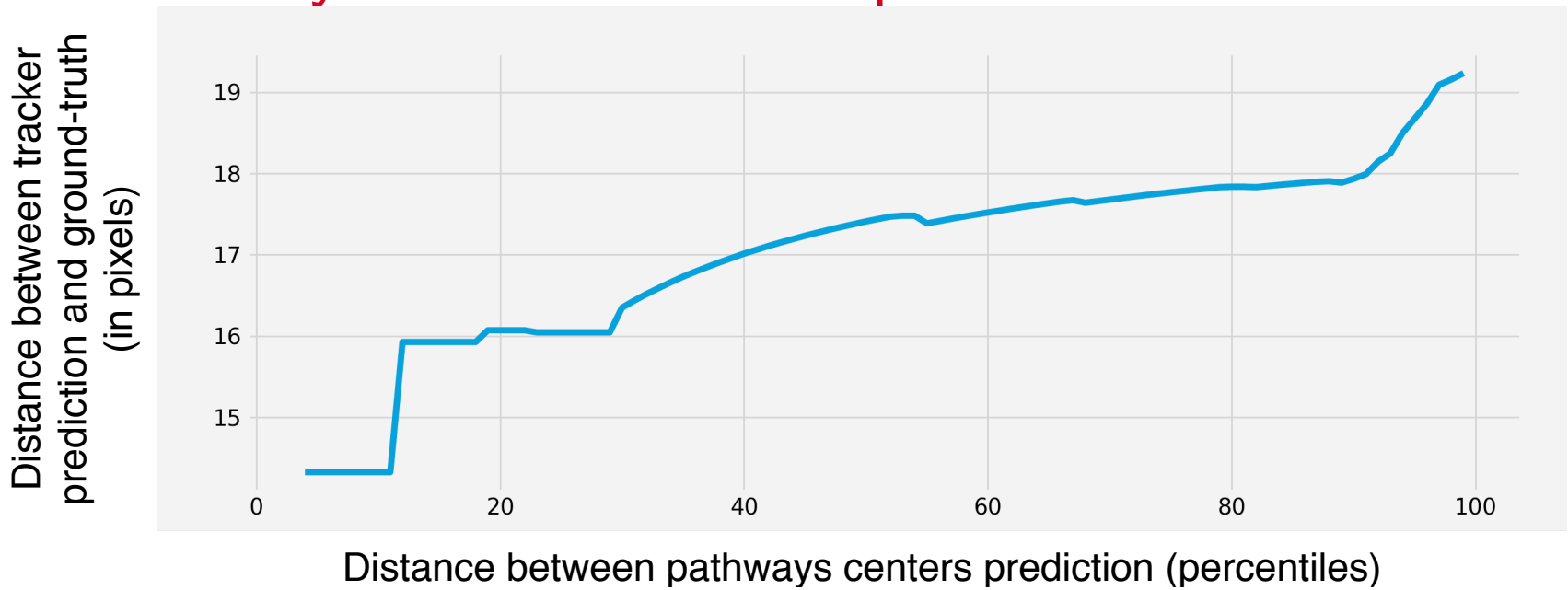
Trackers: He et al. (2017); Danelljan et al. (2017, 2016); Bertinetto et al. (2016); Vojir et al. (2014); Chen et al. (2013); Zhu et al. (2016); Lukezic et al. (2017); Zhang et al. (2017); Cehovin et al. (2016)

Ablation study - Pathways

Version \ Dataset	VOT17			VOT16		
	EAO	Fail rate	Acc	EAO	Fail rate	Acc
FilterParts only	0.25	0.99	0.42	0.306	0.80	0.44
ConvNetPart only	0.205	2.09	0.43	0.265	1.53	0.46
Combined	**0.309**	**0.765**	**0.44**	**0.361**	**0.47**	**0.48**

- "ConvNetPart only": use the FilterParts pathway only for the first 20 frames, to initialize the network
- "ConvNetPart only": the ConvNetPart is trained on each frame, using its own predictions as ground-truth
- ConvNetPart is not stable (high failure rate)

Ablation study - ConvNetPath update



Dataset / Version	VOT17			VOT16		
	EAO	Fail rate	Acc	EAO	Fail rate	Acc
No update	0.28	0.95	0.43	0.34	0.7	**0.48**
Full update	0.284	0.92	**0.44**	0.327	0.66	0.46
HCFs update	**0.309**	**0.765**	**0.44**	**0.361**	**0.47**	**0.48**

- 11% of frames are HCFs

Ablation study - Part roles

Dataset / Version	VOT17			VOT16		
	EAO	Fail rate	Acc	EAO	Fail rate	Acc
One role	0.262	0.99	0.44	0.31	0.715	0.47
All roles	**0.309**	**0.765**	**0.44**	**0.361**	**0.47**	**0.48**

- roles are assigned using spatial and temporal co-occurrences
- roles for FilterParts: candidate, reliable, gold

Conclusions and Future work

- Conclusion
  - 2 deep pathways, co-occurrences constraints in order to keep each path robustness high over time
  - FilterNetPart
    - less flexible, more robust
    - different roles, depending on their degree of reliability
  - ConvNetPart
    - less robust but more capable of adapting to complex changes in object appearance
    - trained only on HCF
  - state of the art results for VOT17 (by large margin for failure rate)
- Future work
  - segmentation for tracker shape
  - speedup for filter parts

Thank you!

References I

Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr. Staple: Complementary learners for real-time tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1401–1409, 2016. doi: 10.1109/CVPR.2016.156. URL https://doi.org/10.1109/CVPR.2016.156.

Luka Cehovin, Ales Leonardis, and Matej Kristan. Robust visual tracking using template anchors. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–8, 2016. doi: 10.1109/WACV.2016.7477570. URL https://doi.org/10.1109/WACV.2016.7477570.

References II

Dapeng Chen, Zejian Yuan, Yang Wu, Geng Zhang, and Nanning Zheng. Constructing adaptive complex cells for robust visual tracking. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1113–1120, 2013. doi: 10.1109/ICCV.2013.142. URL https://doi.org/10.1109/ICCV.2013.142.

Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. 2016.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: efficient convolution operators for tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6931–6939, 2017. doi: 10.1109/CVPR.2017.733. URL https://doi.org/10.1109/CVPR.2017.733.

References III

Sam Hare, Stuart Golodetz, Amir Saffari, Vibhav Vineet, Ming-Ming Cheng, Stephen L. Hicks, and Philip H. S. Torr. Struck: Structured output tracking with kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2096–2109, 2016. doi: 10.1109/TPAMI.2015.2509974. URL http://dx.doi.org/10.1109/TPAMI.2015.2509974.

Zhiqun He, Yingruo Fan, Junfei Zhuang, Yuan Dong, and HongLiang Bai. Correlation filters with weighted convolution responses. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, and Gustavo Fernandez. The visual object tracking vot2017 challenge results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

References IV

Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4847–4856, 2017. doi: 10.1109/CVPR.2017.515. URL https://doi.org/10.1109/CVPR.2017.515.

J. Matas R. Felsberg Pflugfelder M. L. Cehovin G. Vojír T.and Häger M. Kristan, A. Leonardis and et al. The visual object tracking vot2016 challenge results. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 777–823, 2016. doi: 10.1007/978-3-319-48881-3_54. URL https://doi.org/10.1007/978-3-319-48881-3_54.

References V

Kevin P. Murphy. *Machine learning - a probabilistic perspective*. Adaptive computation and machine learning series. MIT Press, 2012. ISBN 0262018020.

Tomas Vojir, Jana Noskova, and Jiri Matas. Robust scale-adaptive mean-shift for tracking. *Pattern Recognition Letters*, 49:250–258, 2014. doi: 10.1016/j.patrec.2014.03.025. URL https://doi.org/10.1016/j.patrec.2014.03.025.

Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Multi-task correlation particle filter for robust object tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4819–4827, 2017. doi: 10.1109/CVPR.2017.512. URL https://doi.org/10.1109/CVPR.2017.512.

References VI

Gao Zhu, Fatih Porikli, and Hongdong Li. Beyond local search: Tracking objects everywhere with instance-specific proposals. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 943–951, 2016. doi: 10.1109/CVPR.2016.108. URL https://doi.org/10.1109/CVPR.2016.108.