PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

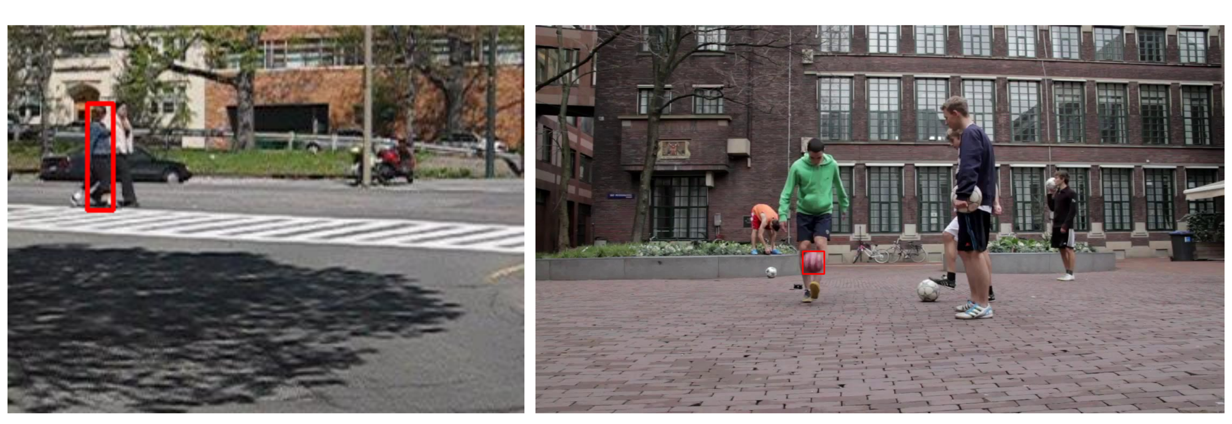
A Simple Adaptive Tracker with Reminiscenses

Christopher Xie[1], Emily Fox[12], Zaid Harchaoui[2]
University of Washington
[1]Paul G. Allen School of Computer Science and Engineering, [2]Department of Statistics

PROBLEM

> Goal: visually track an arbitrary object over time.

> Only a single bounding box in the first frame of the video is given. Examples:

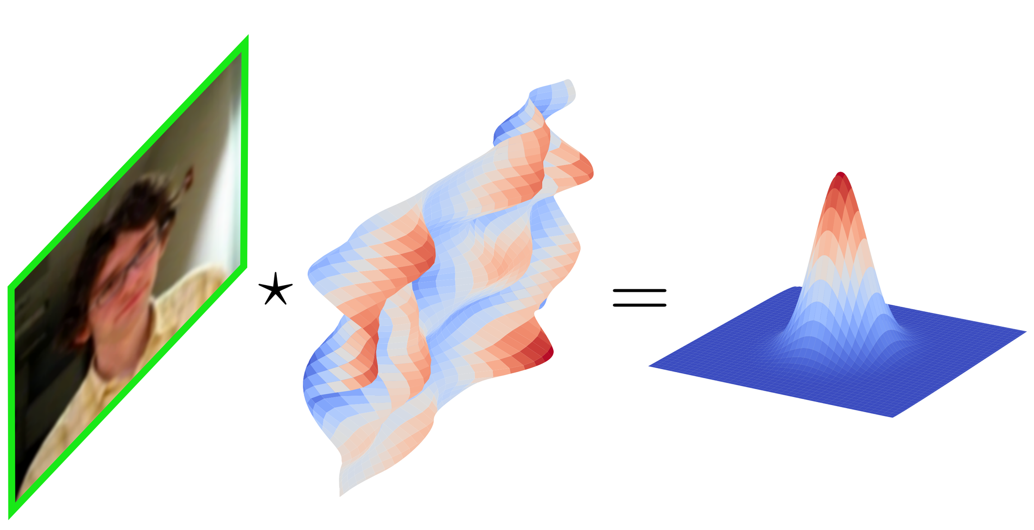


> Difficulties: rotation, scale variation, and object deformation, etc.

Prior Work

> Correlation Filters [1, 2] learn an adaptable object template by minimizing a least squares objective function on Fourier coefficients.

> Issues: inappropriate size due to learning in Fourier domain, learning a single template with short memory.



Approach

> A simple solution that learns a tracker directly in the spatial domain, avoiding known issues while allowing for off-the-shelf gradient-based convex optimization.

> Ensemble-based solution where base trackers are trained on different temporal windows of the video history. Enables robustness to short-term and long-term changes in appearance.

> Our algorithm is denoted the Multi-Template Correlation Filter, or MTCF.

References

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

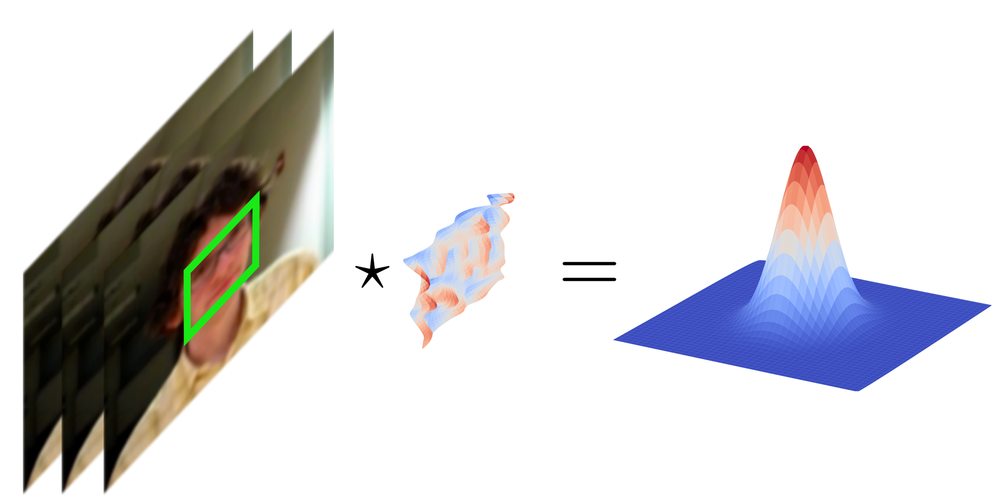
[2] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

METHOD

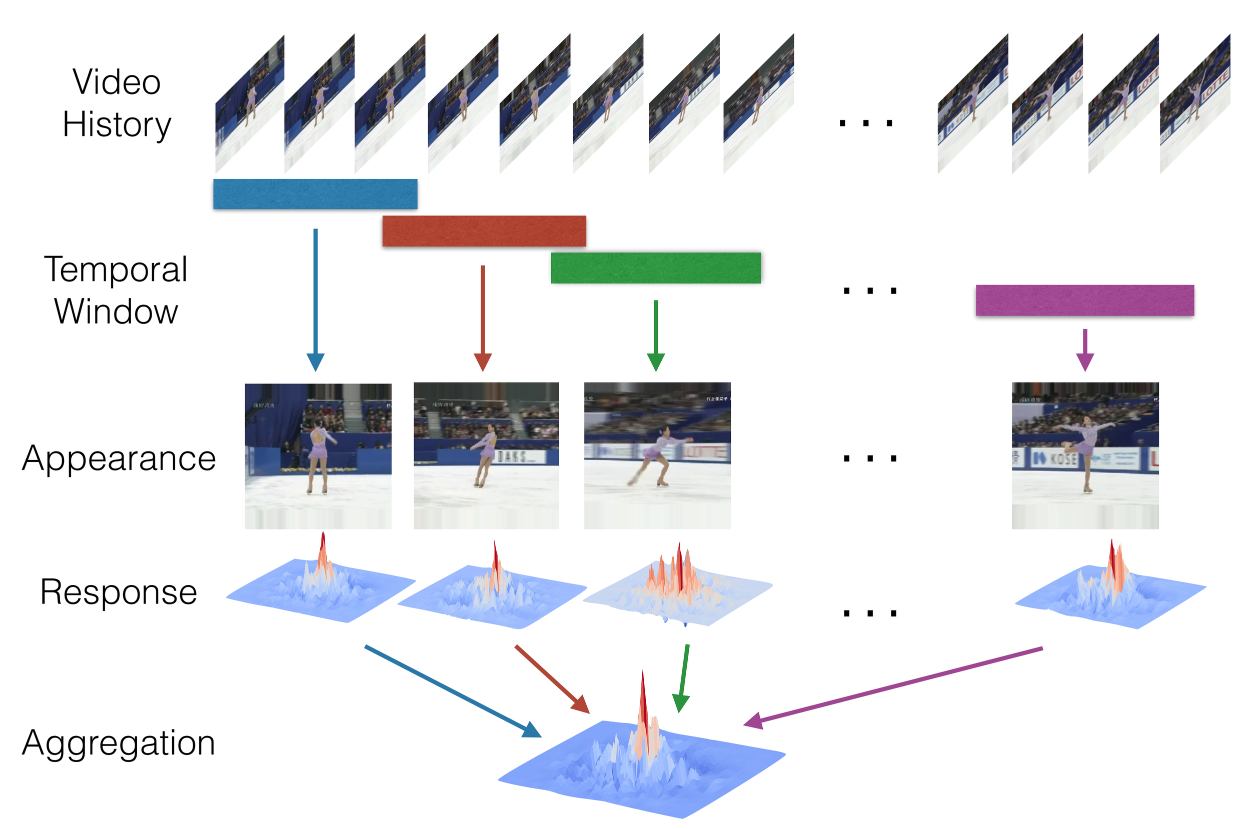
Base Tracker

$$F^* = \underset{F}{\operatorname{argmin}} \frac{1}{2} \sum_{t=1}^{N} \alpha_t \left\| Y_t - \sum_{k=1}^{d} [F]_k \star [I_t]_k \right\|_2^2 + \frac{\lambda}{2} \|F\|_2^2$$

> Visual representation of the model:



MTCF

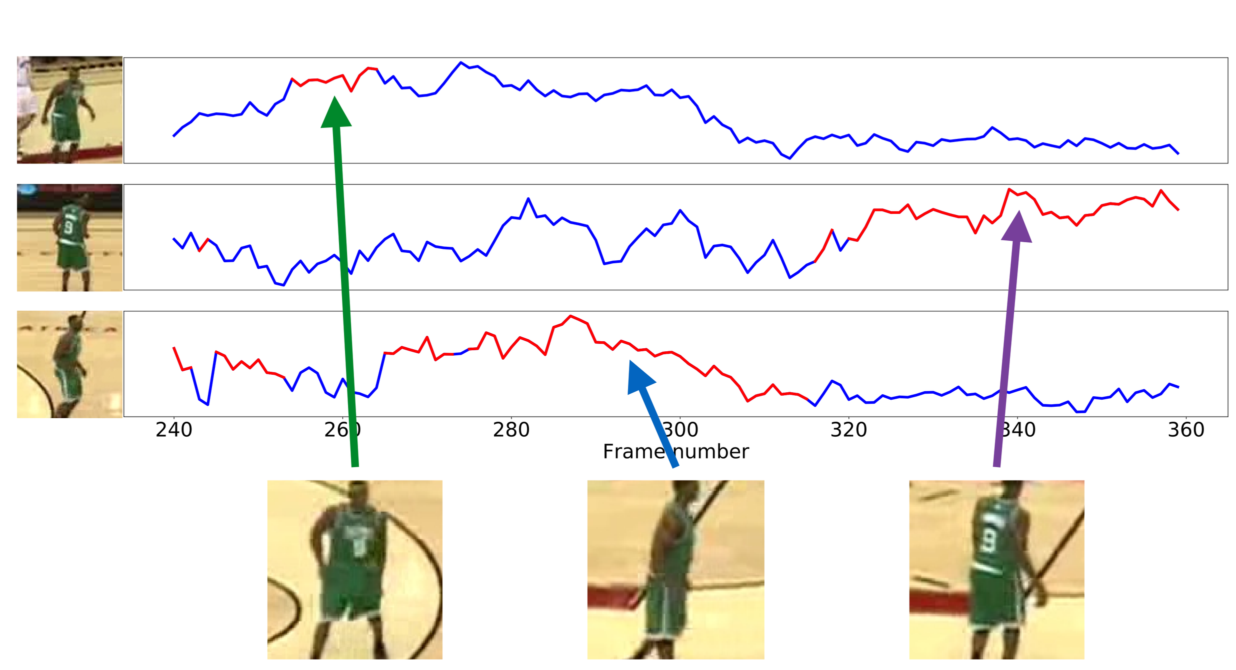


> Aggregated heatmap equation:

$$M = \sum_{i=1}^{L} w_i M_i, \quad w_i = \frac{|D_i|(1-\gamma)^{L-i}}{\sum_{j=1}^{L} |D_j|(1-\gamma)^{L-j}}$$

Demonstration

> Each row shows a different base tracker's per-frame confidence and appearance model.
> Red portions indicate the highest confidence.
> Around frame 260, we see that the object appearance indeed is similar to that of base tracker 1.



EXPERIMENTS

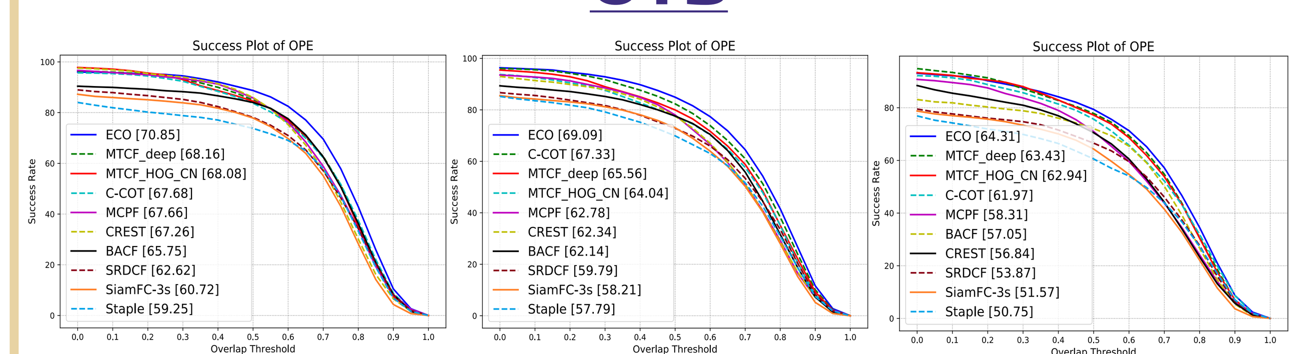
Model Analysis

	OTB-2013	OTB-100	OTB-50	FPS
SRDCF [2]	62.6/78.1	59.8/72.8	53.9/66.6	4.3
sCF - HOG	63.0/80.6	58.6/71.4	53.5/65.4	9.8
MTCF - HOG	**66.0/84.1**	**62.7/77.5**	**59.0/73.2**	9.6
sCF - HOG+CN	63.9/79.5	62.1/75.1	59.2/72.9	8.6
MTCF - HOG+CN	**68.1/84.5**	**64.0/77.5**	**62.9/77.2**	7.3
sCF - deep	67.0/83.1	65.5/79.6	62.0/75.3	2.8
MTCF - deep	**68.2/85.0**	**65.6/80.0**	**63.4/77.8**	2.7

Table 1: AUC and success rates are shown for each of the models.

> Temporal ensemble boosts performance
> Deep network features boosts performance
> Comparable speed to state of the art

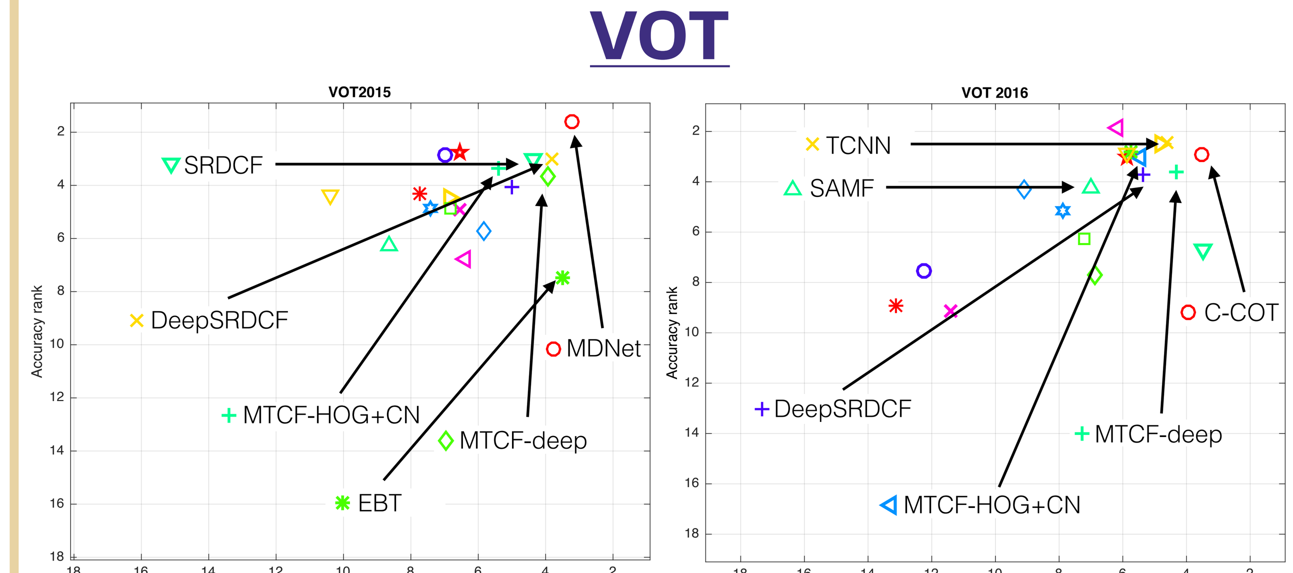
OTB



OTB-2013 OTB-100 OTB-50

> MTCF outperforms almost all SOTA trackers
> HOG+CN features perform quite strongly

VOT



VOT-2015 VOT-2016

> Competitive performance with winning trackers in both years of challenges

Qualitative Examples