Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking


Martin Danelljan


Andreas Robinson


Fahad Khan


Michael Felsberg

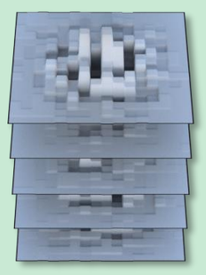
LINKÖPING UNIVERSITY

Computer Vision Laboratory, Linköping University, Sweden

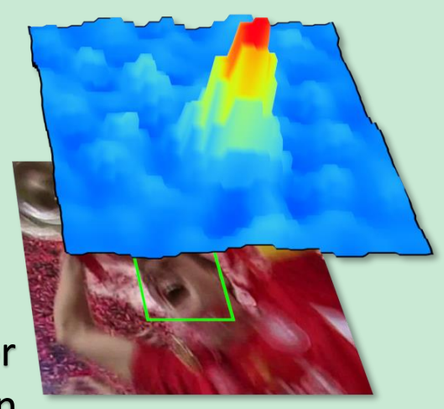
Introduction

Discriminative Correlation Filters (DCF):

Single-resolution feature map



Learns a set of discrete filters for target localization



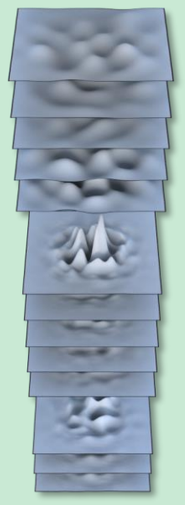
Outputs discrete detection scores

Our Approach:

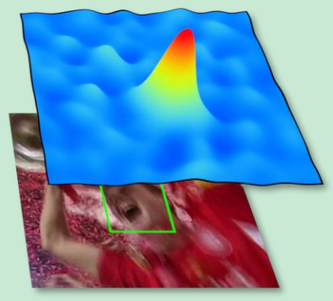
Posing the learning problem in the continuous spatial domain

Multi-resolution (deep) feature map

Learns continuous filters



Outputs continuous detection scores



Advantages

- Integration of multi-resolution (deep) features
- Accurate sub-pixel (or sub-grid) localization
- Sub-pixel supervision in the learning
- Efficient processing of all available information
- Avoids artefacts caused by explicit resampling

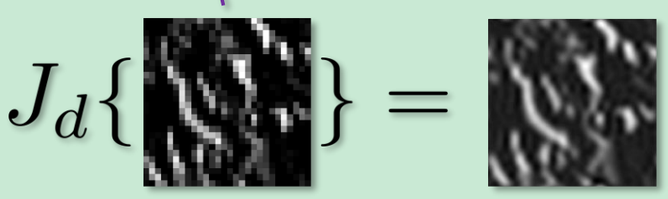
Applications

1) Object tracking 2) Feature point tracking

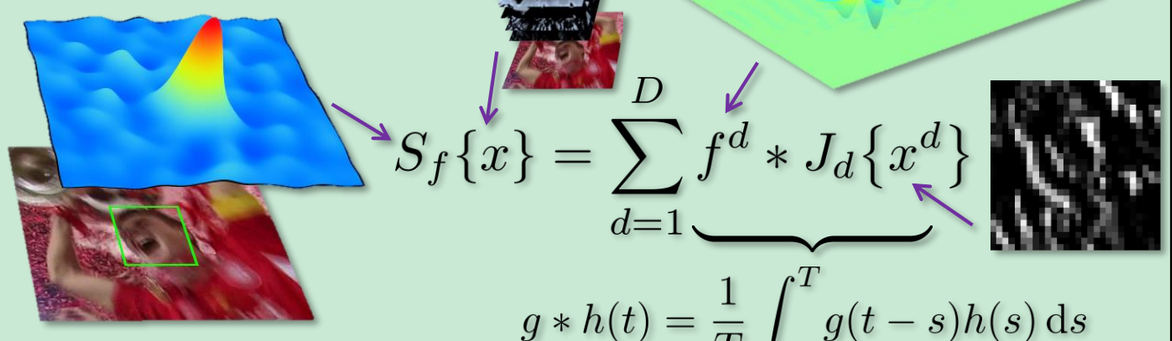
Continuous Convolution Operators

Interpolation Operator $J_d : \mathbb{R}^{N_d} \to L^2(T)$

$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d \left(t - \frac{T}{N_d} n \right)$$



Convolution Operator



$$S_f\{x\} = \sum_{d=1}^{D} f^d * J_d\{x^d\}$$

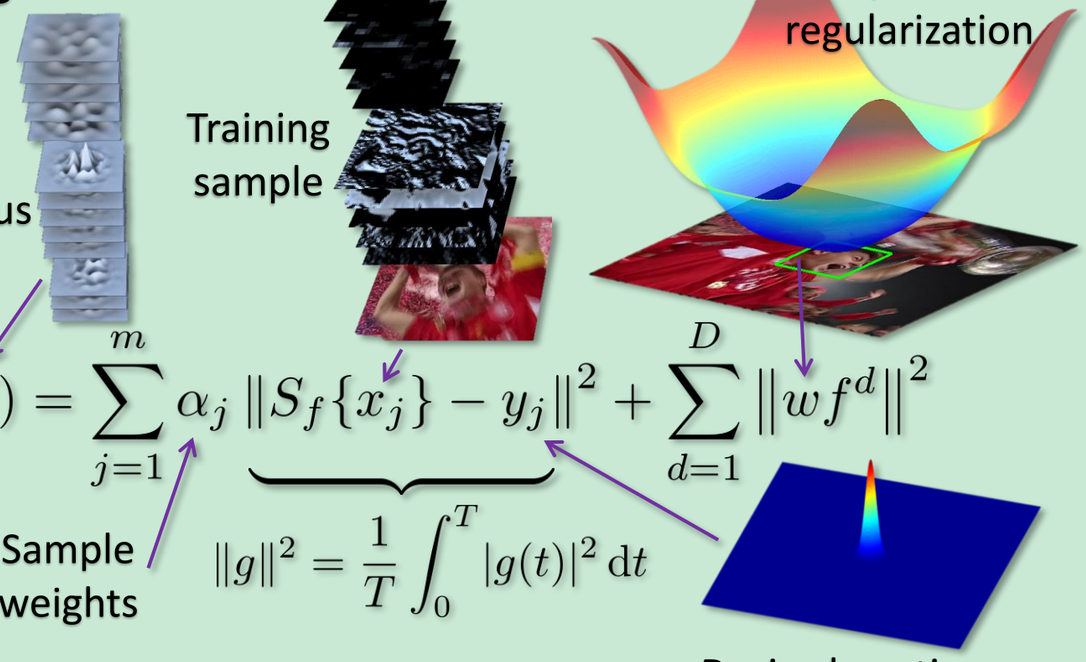
$$g * h(t) = \frac{1}{T} \int_0^T g(t-s) h(s) \, \mathrm{d}s$$

Notation

- $\hat{g}[k]$ - Fourier coefficients of $g \in L^2(T)$
- $X_j^d[k]$ - discrete Fourier transform of $x_j^d \in \mathbb{R}^{N_d}$

Convolution Operator Learning

Training loss



$$E(f) = \sum_{j=1}^{m} \alpha_j \underbrace{\left\| S_f\{x_j\} - y_j \right\|^2}_{\|g\|^2 = \frac{1}{T}\int_0^T |g(t)|^2 \, \mathrm{d}t} + \sum_{d=1}^{D} \left\| w f^d \right\|^2$$

Fourier Domain

$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| \sum_{d=1}^{D} \hat{f}^d X_j^d \hat{b}_d - \hat{y}_j \right\|_{\ell^2}^2 + \sum_{d=1}^{D} \left\| \hat{w} * \hat{f}^d \right\|_{\ell^2}^2$$

Assumption: finitely many non-zero Fourier coefficients.
Gives normal equations: $\left(A^{\mathrm{H}} \Gamma A + W^{\mathrm{H}} W \right) \hat{\mathbf{f}} = A^{\mathrm{H}} \Gamma \hat{\mathbf{y}}$

Object Tracking Framework

- Features: VGG network (pre-trained on ImageNet)
- Optimization: Conjugate Gradient

Feature Point Tracking Framework

Grayscale pixel features $D = 1$

$$\hat{f}[k] = \frac{\sum_{j=1}^{m} \alpha_j \overline{X_j[k] \hat{b}[k]} \hat{y}_j[k]}{\sum_{j=1}^{m} \alpha_j \left| X_j[k] \hat{b}[k] \right|^2 + \beta^2}$$

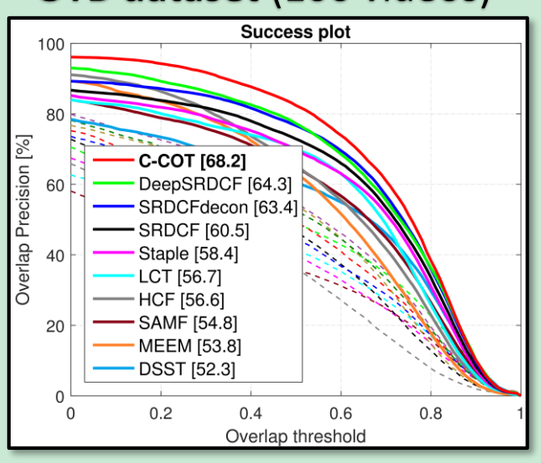
Uniform regularization $w(t) = \beta$

Experiments

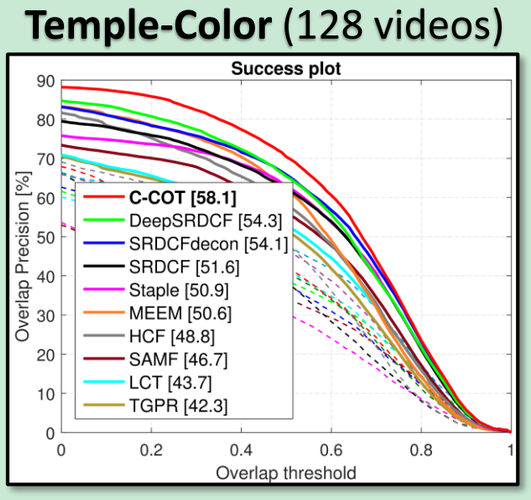
Object Tracking: Layer fusion on OTB (100 videos)

	Layer 0	Layer 1	Layer 5	Layers 0, 1	Layers 0, 5	Layers 1, 5	Layers 0, 1, 5
Mean OP	58.8	78.0	60.0	77.8	70.7	*81.8*	**82.4**
AUC	49.9	65.8	51.1	65.7	59.0	*67.8*	**68.2**

OTB dataset (100 videos)



Temple-Color (128 videos)



VOT2016 challenge results (top 3) [Matej et al., VOT workshop 2016]

	Tracker	EAO	A	R	A_{rank}	R_{rank}	AO	EFO	Impl.
1.	C-COT	**0.331**	0.539	*0.238*	12.000	**1.000**	0.469	0.507	D M
2.	TCNN	*0.325*	0.554	0.268	4.000	*2.000*	*0.485*	1.049	S M
3.	SSAT	0.321	**0.577**	0.291	**1.000**	3.000	**0.515**	0.475	S M

Feature Point Tracking: The Sintel dataset