Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking

Martin Danelljan, Andreas Robinson,
Fahad Shahbaz Khan, Michael Felsberg

LINKÖPING UNIVERSITY

Discriminative Correlation Filters (DCF)

Applications

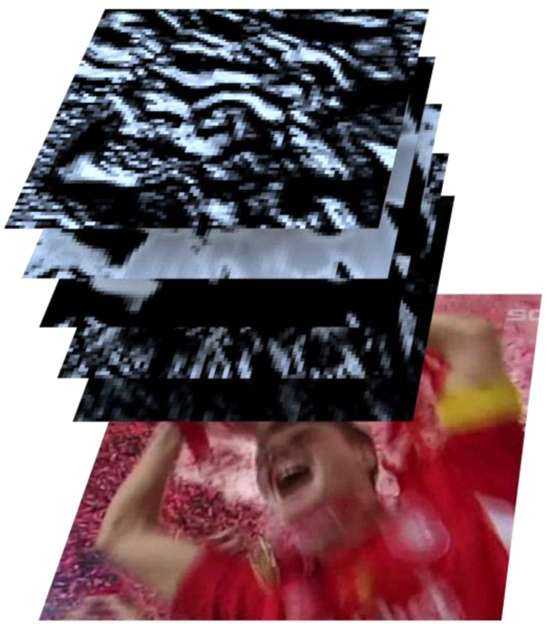
- Object recognition
- Object detection
- Object tracking
  - Among state-of-the-art since 2014
  - KCF, DSST, HCF, SRDCF, Staple …

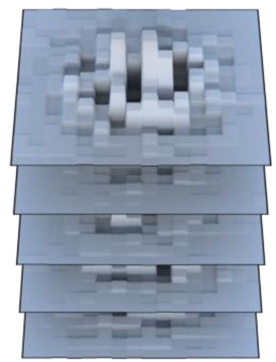
Discriminative Correlation Filters (DCF)

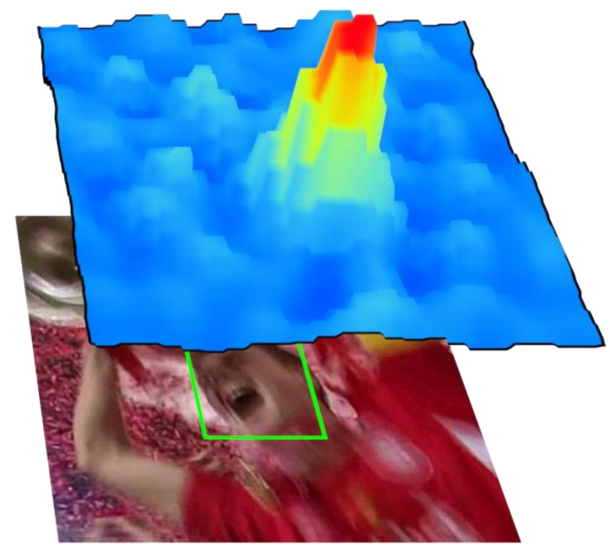
Limitations:

Single-resolution feature map

Coarse output scores







DCF Limitations:
1. Single-resolution feature map

- Why a problem?
  - Combine convolutional layers of a CNN
    - Shallow layers: low invariance – high resolution
    - Deep layers: high invariance – low resolution
- How to solve?
  - Explicit resampling?
    - Artefacts, information loss, redundant data
  - Independent DCFs with late fusion?
    - Sub-optimal, correlations between layers

DCF Limitations:
2. Coarse output scores

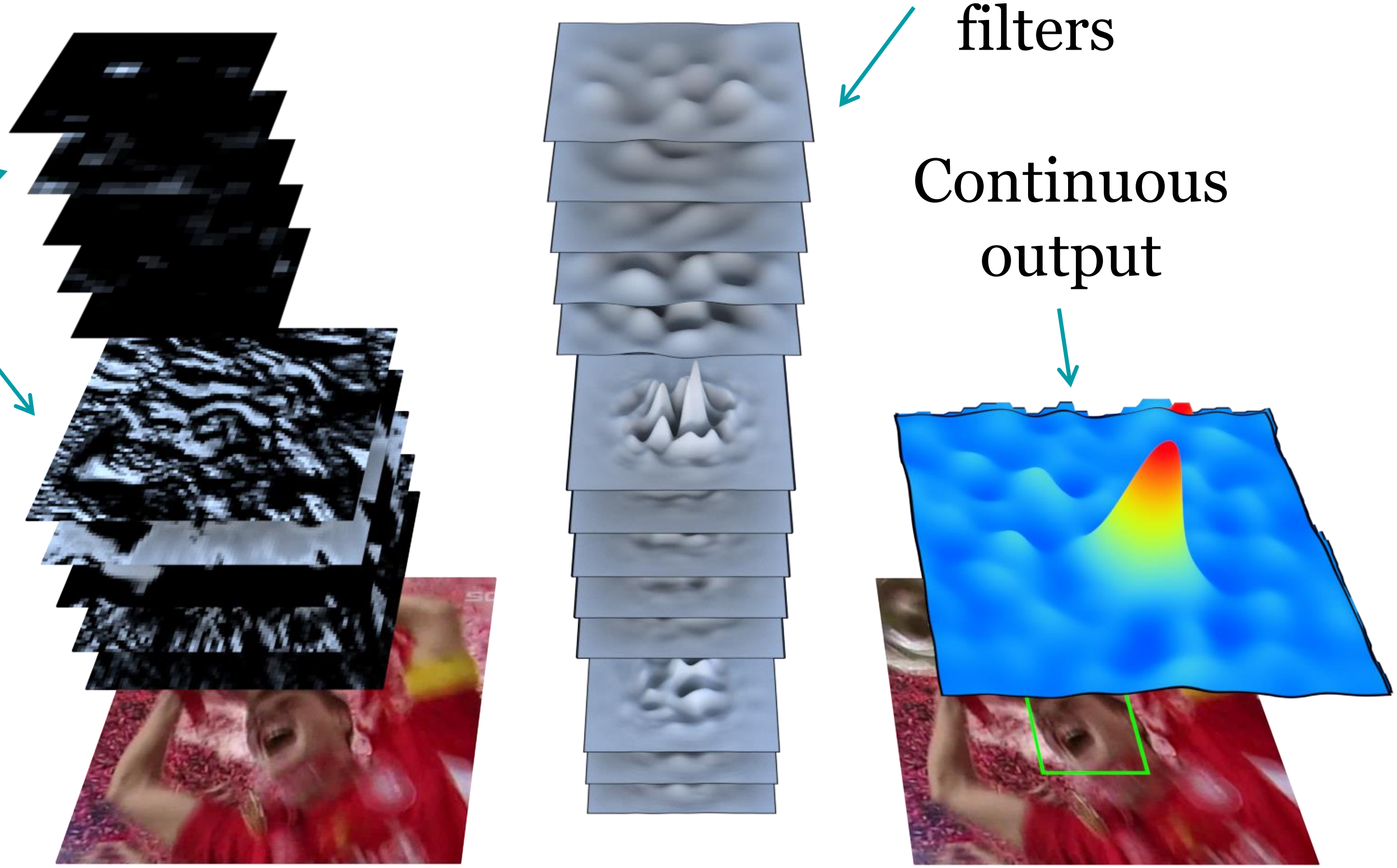
- Why a problem?
  - Accurate localization
    - Sub-grid (e.g. HOG grid) or sub-pixel accuracy
    - More accurate annotations=> less drift
- How to solve?
  - Interpolation?
    - Which interpolation strategy?
  - Interweaving?
    - Costly

DCF Limitations:

3. Coarse labels

- Why a problem?
  - Accurate learning
    - Sub-grid or sub-pixel supervision
- How to solve?
  - Interweaving?
    - Costly
  - Explicit interpolation of features?
    - Artefacts

Our Approach: Overview



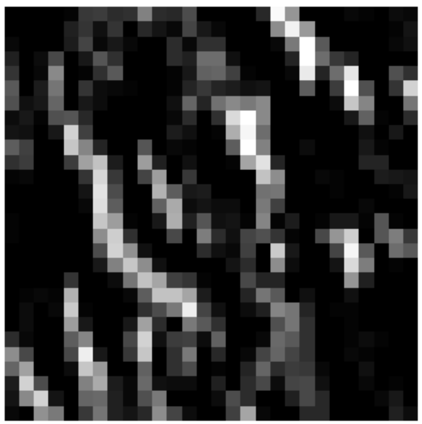
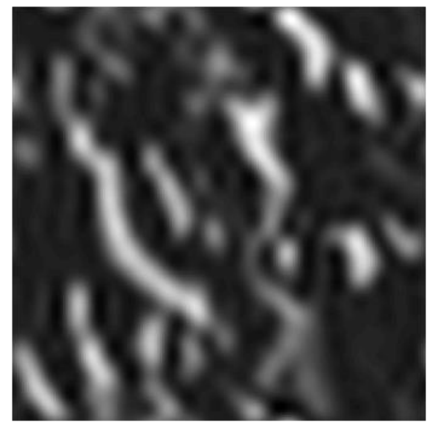
Multiresolution Features

$$x = \{$$

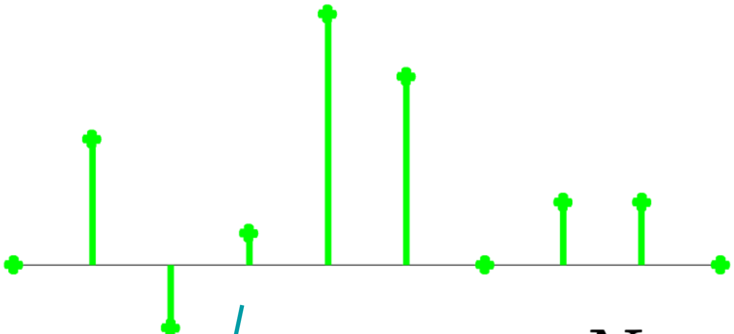
$$x^d \in \mathbb{R}^{N_d}$$

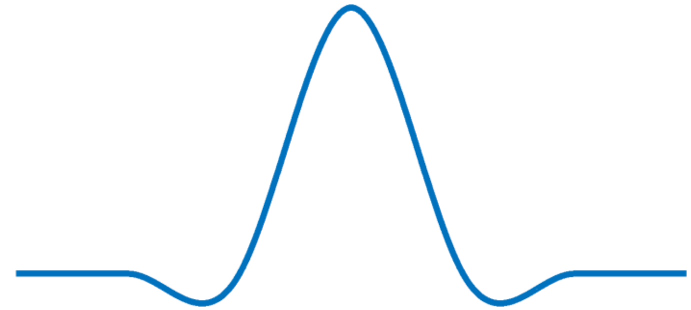
Interpolation Operator

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

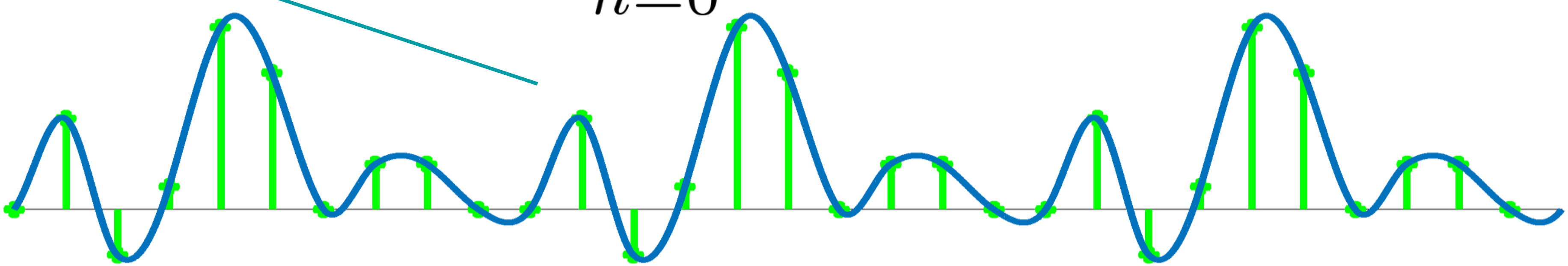
$J_d\{$  $\} =$ 

Interpolation Operator

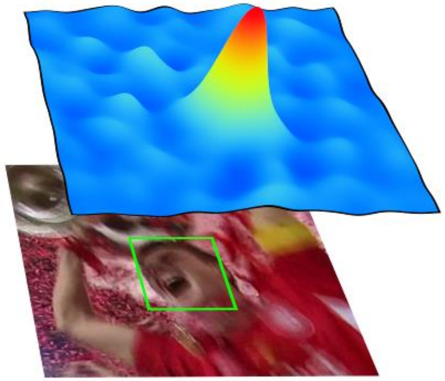


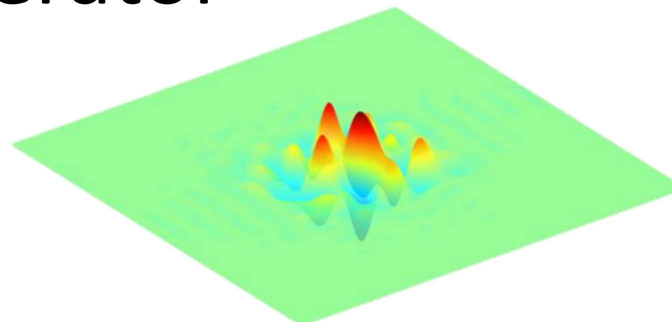


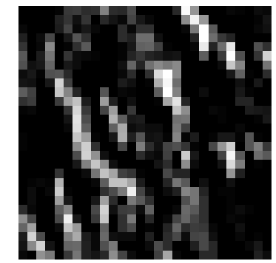
$$J_d\{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d\left(t - \frac{T}{N_d} n\right)$$



Convolution Operator



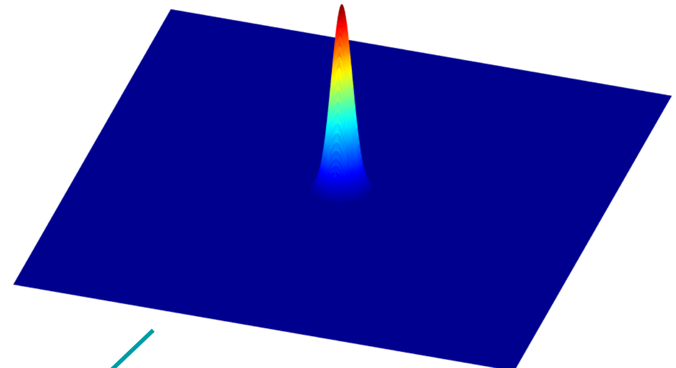




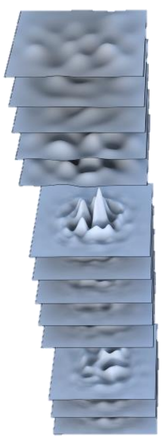
$$S_f\{x\} = \sum_{d=1}^{D} \underbrace{f^d * J_d\{x^d\}}$$

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

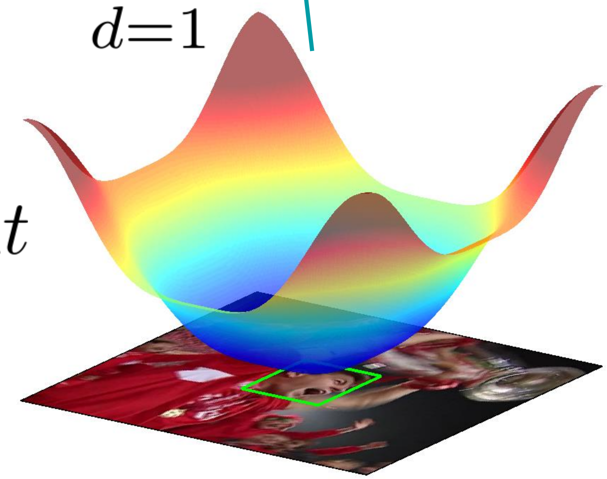
Training Loss



$$E(f) = \sum_{j=1}^{m} \alpha_j \left\| S_f\{x_j\} - y_j \right\|^2 + \sum_{d=1}^{D} \left\| w f^d \right\|^2$$



$$\|g\|^2 = \frac{1}{T} \int_0^T |g(t)|^2 \, \mathrm{d}t$$



[Danelljan et al., ICCV 2015]

Training Loss – Fourier Domain

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

$$\hat{g}[k] = \langle g, e_k \rangle = \frac{1}{T} \int_0^T g(t) e^{-i \frac{2\pi}{T} kt} \, \mathrm{d}t$$

$$X^d[k] = \sum_{n=0}^{N_d - 1} x^d[n] e^{-i \frac{2\pi}{N_d} nk}$$

Training Loss – Fourier Domain

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

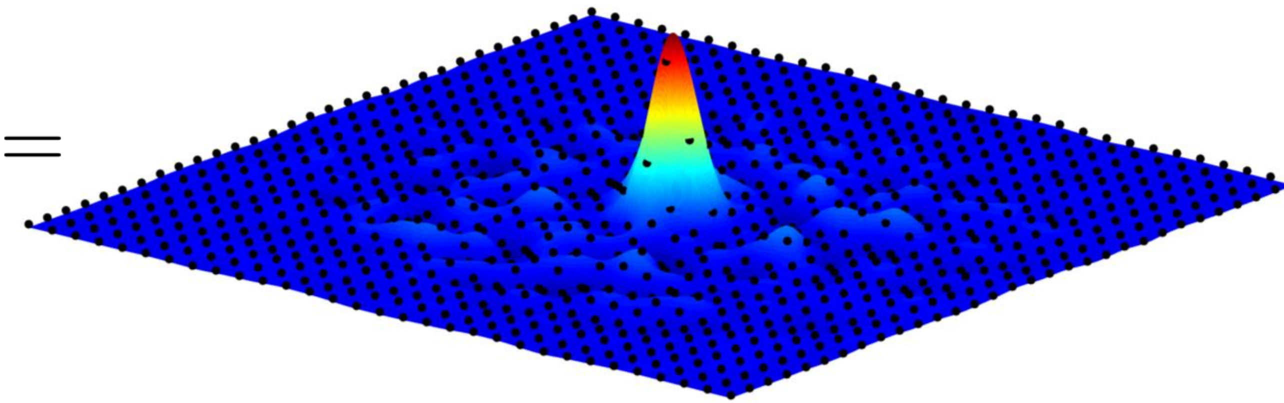
$$\Downarrow$$

$$\left(A^{\mathrm{H}} \Gamma A + W^{\mathrm{H}} W\right) \hat{\mathbf{f}} = A^{\mathrm{H}} \Gamma \hat{\mathbf{y}}$$

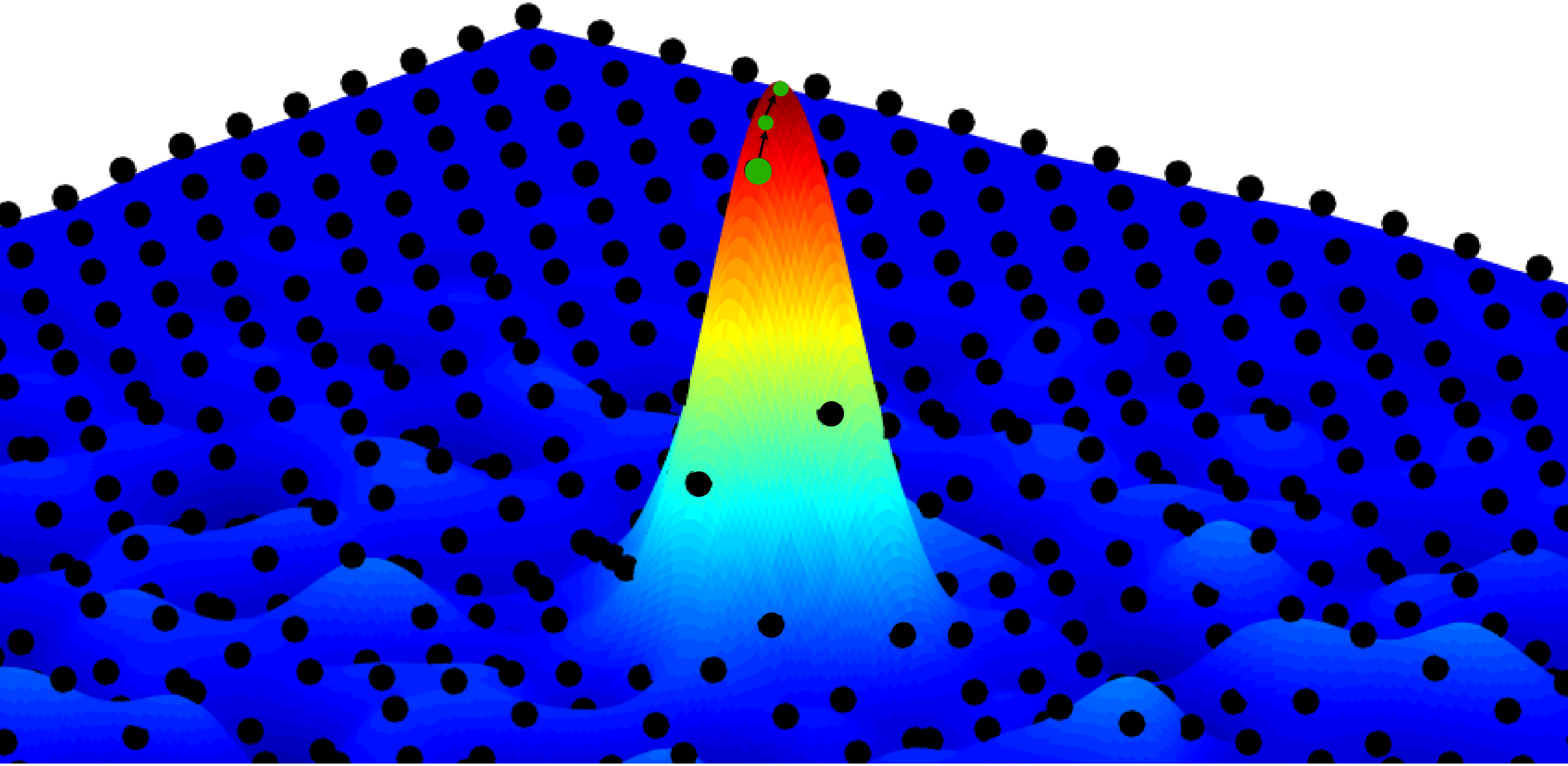
Localization

$S_f\{$

$\} =$



Localization



How to set y_j and b_d ?

- Use periodic summation of functions $g : \mathbb{R} \to \mathbb{R}$:

$$g_T(t) = \sum_{n=-\infty}^{\infty} g(t - nT)$$

- Gaussian function for y_j
- Cubic spline kernel for b_d
- Fourier coefficients $\hat{y}_j, \hat{b}_d$ with Poisson's summation formula:

$$\hat{g}_T[k] = \frac{1}{T}\hat{g}(\frac{k}{T})$$

Object Tracking Framework: Features

- VGG network
  - Pre-trained on ImageNet
  - No fine-tuning on application specific data

Object Tracking Framework: Optimization

Solving $\left(A^{\mathrm{H}}\Gamma A + W^{\mathrm{H}}W\right)\hat{\mathbf{f}} = A^{\mathrm{H}}\Gamma\hat{\mathbf{y}}$

SRDCF: Gauss-Seidel

- ☹ Explicit computation of $\left(A^{\mathrm{H}}\Gamma A + W^{\mathrm{H}}W\right)$
- 😐 Sparse matrix handling
- ☹ $\mathcal{O}(D^2)$
- 😐 "Infinite" memory
- ☺ Warm starting: trivial

C-COT: Conjugate Gradient

- ☺ Only need to evaluate $\left(A^{\mathrm{H}}\Gamma A + W^{\mathrm{H}}W\right)\hat{\mathbf{f}}$
- ☺ **No** sparse matrix handling
- ☺ $\mathcal{O}(D)$
- 😐 Finite memory
- 😐 Warm starting: **non**-trivial
- 😐 Tuning of pre-conditioners

Object Tracking Framework: Pipeline

- Simple:

  … – Track – Train – Track – Train – …

- No thresholds
- No hidden "tricks"

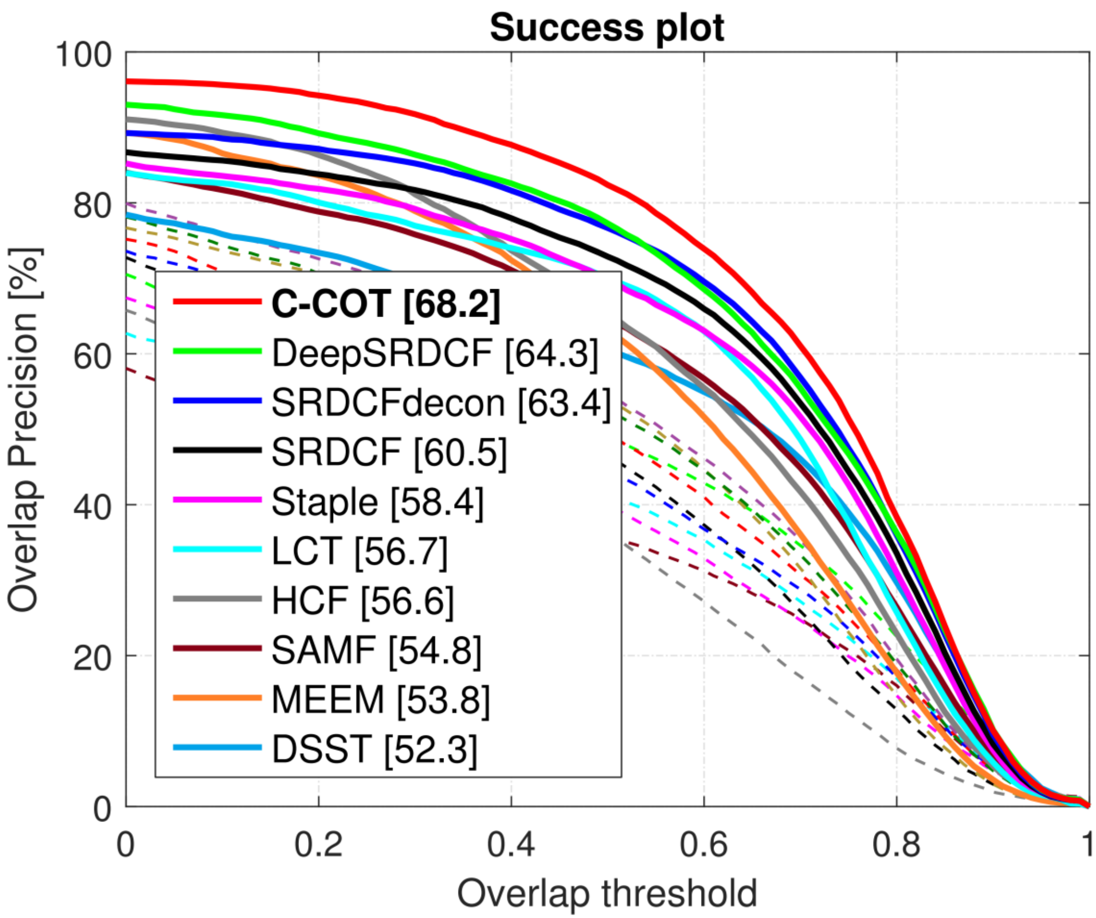
Experiments: Object Tracking

- 3 datasets: OTB-100, TempleColor, VOT2015
- Layer fusion on OTB:

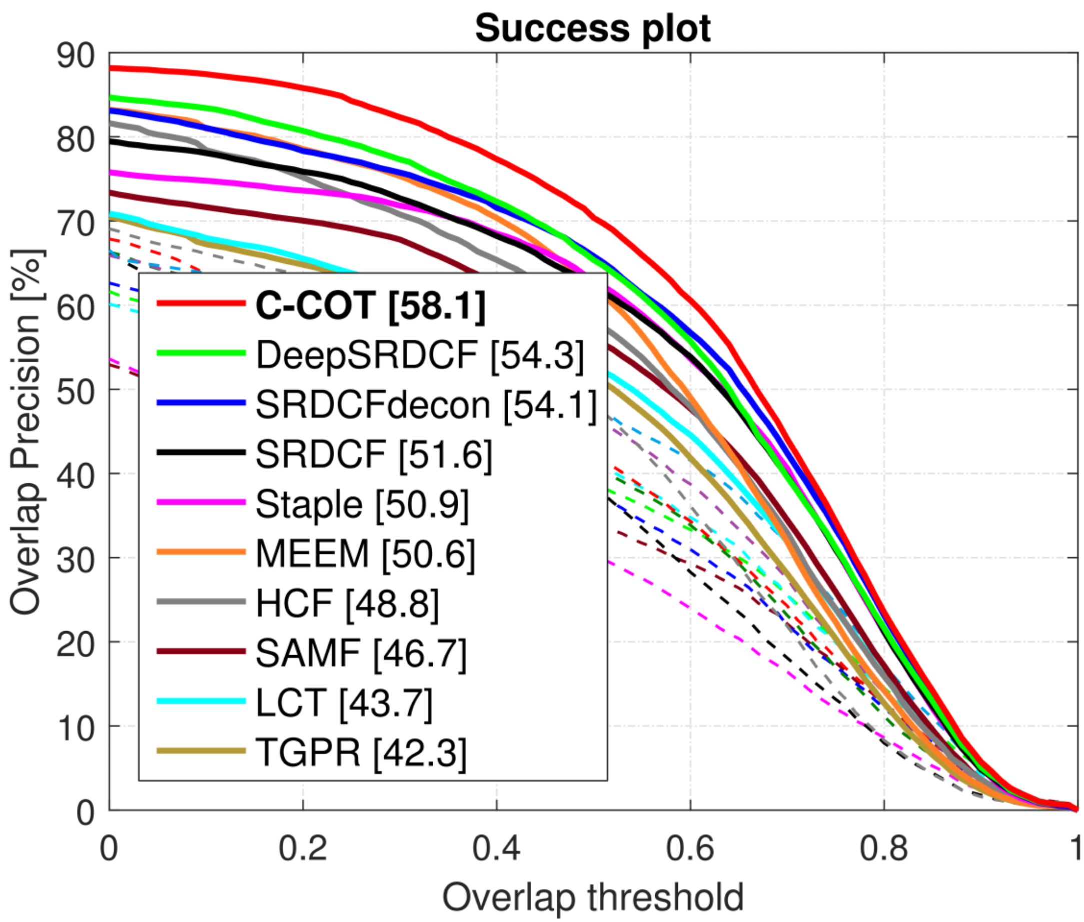
	Layer 0	Layer 1	Layer 5	Layers 0, 1	Layers 0, 5	Layers 1, 5	Layers 0, 1, 5
Mean OP	58.8	78.0	60.0	77.8	70.7	*81.8*	**82.4**
AUC	49.9	65.8	51.1	65.7	59.0	*67.8*	**68.2**

- Compared to explicit resampling in DCF
  - Performance gain: $+7.4\%$ AUC
  - Efficiency gain: -80% data size

Experiments: OTB (100 videos)



Experiments: Temple-Color (128 videos)



Experiments: VOT2016

	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
4.	MLDF	0.311	0.490	**0.233**	36.000	**1.000**	0.428	1.483	D M
5.	Staple	0.295	0.544	0.378	5.000	10.000	0.388	11.144	D C

[Matej et al., ECCV VOT workshop 2016]

Object Tracking: Speed

- With CNN features: slow ~1 FPS (no GPU)
- With HOG features: ~ real time at SRDCF performance

Feature Point Tracking Framework

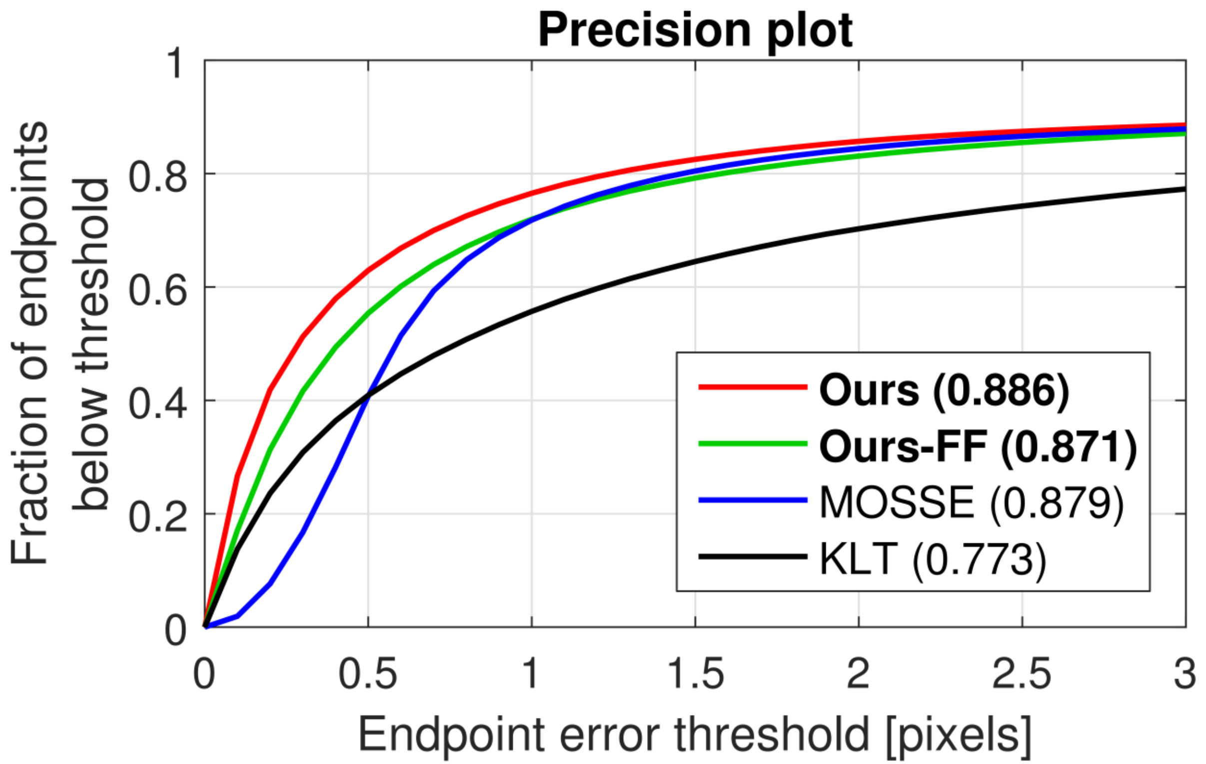
- Grayscale pixel features, $D = 1$
- Uniform regularization, $w(t) = \beta$

$$\Downarrow$$

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

Experiments: Feature Point Tracking

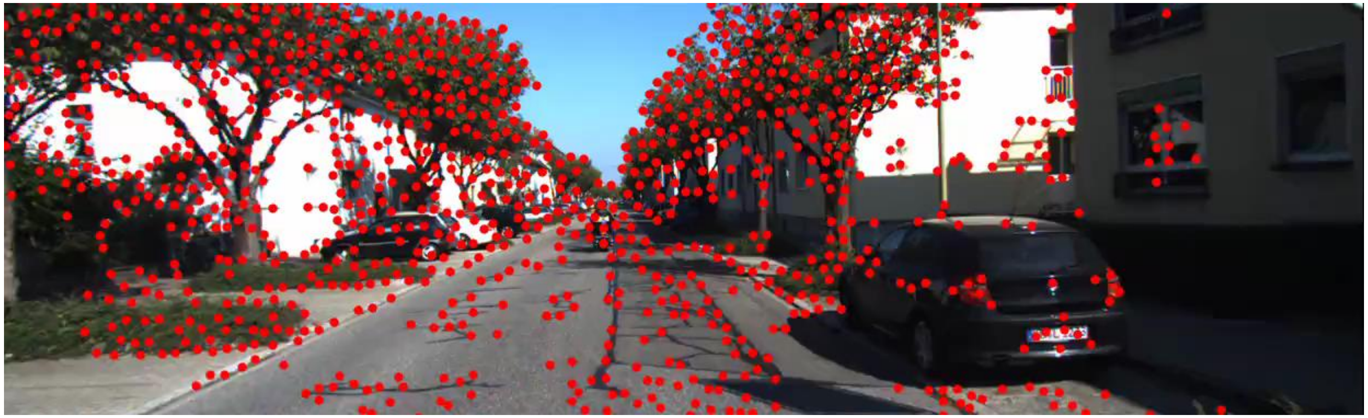
- The Sintel dataset



Feature Point Tracking: KITTI

C-COT

KLT



Future Work

- Features
  - Fine tuning
  - Unsupervised learning
- Optimization
  - Warm start in CG (theory and heuristics)
  - Preconditioners
  - Implementation aspects
  - Alternative strategies or update rules
- Further explore of the continuous formulation

Conclusions

- **Continuous domain** learning formulation
  - **Multi-resolution** deep feature maps
  - **Sub-pixel** accurate localization
  - **Sub-pixel** supervision
- Superior results for two applications
  - Object tracking
  - Feature point tracking

Oral and poster: O-4B-03

Friday afternoon
(last session)

Martin Danelljan

www.liu.se

LINKÖPING
UNIVERSITY