

# Kernalised Multi-resolution Convnet for Visual Tracking

Di Wu, Wenbin Zou,\* Xia Li  
Shenzhen University<sup>†</sup>  
dwu, wzou, lixia@szu.edu.cn

Yong Zhao  
Peking University Shenzhen Graduate School  
yongzhao@pkusz.edu.cn

## Abstract

Visual tracking is intrinsically a temporal problem. Discriminative Correlation Filters (DCF) have demonstrated excellent performance for high-speed generic visual object tracking. Built upon their seminal work, there has been a plethora of recent improvements relying on convolutional neural network (CNN) pretrained on ImageNet as a feature extractor for visual tracking. However, most of their works relying on ad hoc analysis to design the weights for different layers either using boosting or hedging techniques as an ensemble tracker. In this paper, we go beyond the conventional DCF framework and propose a Kernalised Multi-resolution Convnet (KMC) formulation that utilises hierarchical response maps to directly output the target movement. When directly deployed the learnt network to predict the unseen challenging UAV tracking dataset without any weight adjustment, the proposed model consistently achieves excellent tracking performance. Moreover, the transfered multi-resolution CNN renders it possible to be integrated into the RNN temporal learning framework, therefore opening the door on the end-to-end temporal deep learning (TDL) for visual tracking.

## 1. Introduction

Visual tracking is the task of predicting the trajectory of a target in a video. The task of tracking, a crucial component of many computer vision systems, can be naturally specified as an online learning problem. This paper focuses on the challenging problem of monocular, generic, realistic object tracking.

Visual object tracking has recently witnessed substantial progress due to powerful features extracted using deep convolutional neural networks. For online applications, one simple approach to transfer offline pre-trained CNN features is to add one or more randomly initialised CNN layers, named as adaptation layers, on top of the pre-trained CNN

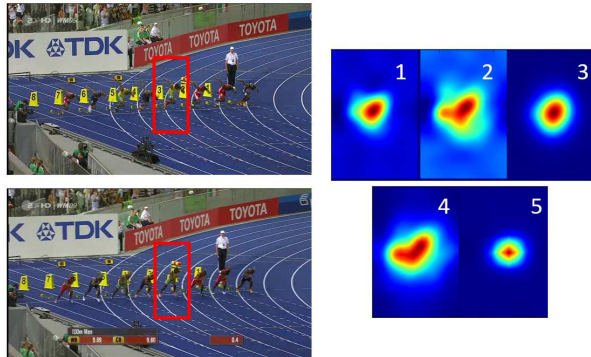


Figure 1: Response maps generated from hierarchical deep features. Left: two frames from the “Bolt” sequence, red boxes are image patches with padding added. Right: hierarchical response maps from correlation filters (response maps have been resized and rescaled to the same shape and scale). It can be seen that because different layers of deep features encode different semantics, the maximum response points from different layers also vary.

model. However, as [26] empirically observe that this transfer learning method suffers from severe over-fitting because of the limited training data. It is also observed in [29, 30] that the differences in mean activations in intermediate layers have noticeable effect for multi-model fusion. The online learnt parameters mainly focus on recent training samples and are less likely to be well generalised to both historical and future samples. This phenomenon can be fatal to online visual tracking where the target often undergoes significant appearance changes or heavy occlusion. CNNs also have shown impressive performance as an offline feature extractor for tracking [27, 18] in lieu of traditional handcrafted features (*e.g.*, HOG, Color Moment). Features from these deep convolutional layers are discriminative while preserving spatial and structural information.

The marriage between DCF [9], which has the advantage of being efficient in training translational images in the fourier space, and deep features, which excel at image representation, further advances the visual tracking community [21, 7]. However, there are some conflicting con-

\*Corresponding author.

<sup>†</sup>Shenzhen Key Lab of Advanced Telecommunication and Information Processing, College of Information Engineering, Shenzhen University.

clusions concerning the representation power from different layers of CNN: shallow convolutional layers have been highlighted in the DCF-based methods [5] whereas [21] found that the performance generally increases as layer depth is increased hence only convolutional layers deeper than 10th is used. Fig. 1 shows the response maps generated from hierarchical deep features. Since different layers from convnet capture different semantic meanings, maximum translational locations obtained via different convnet layers could vary: low level layers excel at discriminating intra-class variations whereas high level layers excel at discriminating inter-class variations. Hence, for generic object tracking, it is imperative to combine features from different layers to have a generalised representation from hierarchies of response maps.

Visual tracking is intrinsically also a temporal problem, but many previous approaches [1, 26] mainly focus on the design of a robust appearance model. An end-to-end object tracking approach which directly maps from raw sensor input to object tracks in sensor space is proposed in [20]. Their proposed method works well in a simulated environment, but the applicability on the realistic RGB imagery dataset is worth further investigation. One reason that temporal deep learning based techniques are challenging to be employed in realistic settings is due to the difficulty for recurrent net to have robust, meaningful temporal input.

In this paper, we propose a kernelised multi-resolution convnet tracking algorithm that utilises the intermediate response maps from the correlation filter output and learns the implicit translational output from the multi-resolution convnet. The key contributions can be summarised as follows:

- We incorporate kernelised form into CNN feature representations for the non-linear regression tasks that consistently outperforms its linear counterpart, concluding the kernelised version should be the preferred choice in the process of designing correlation filters.
- We learn a novel representation for multi-resolution response maps generated from different layers of a pre-trained CNN, negating the need to design the weights for various convnets layers. Moreover, the learnt representation renders it possible to be included into the end-to-end temporal deep learning (TDL) pipeline.
- We use an adaptive hedge method to update the model learning rate, taking the model stability into consideration, making the update of target model smoother and more robust.

## 2. Related Work

We review recent tracking methods closely related to this work as in the following three sections.

**Correlation filter based trackers:** Most recent works [1, 26] strives for a better appearance models for the tracker. DCF [1, 9] have demonstrated excellent performance for high-speed visual object tracking. The key to their success is by observing that the resulting data matrix of translated patches is circulant, the cyclic shifts could be diagonalised with the Discrete Fourier Transform, reducing both storage and computation to obtain the next frame response map. However, most of these trackers depend on the spatiotemporal consistency of visual cues. Therefore, they can handle mostly short-term tracking or object with static appearance.

**Deep feature based trackers:** Recent works exploit the structure of CNN to learn the target online: a three-layer CNN is trained on-the-flight in [15]; a deep autoencoder [28] is first pre-trained offline and then finetuned for binary classification in online tracking. Since the pre-training is performed in an unsupervised way by reconstructing gray images with very low resolution, the learned deep features has limited discriminative power for tracking. Moreover, without pre-training and with limited training samples obtained online, CNN fails to capture object semantics and is not robust to deformation. Both [15] and [28] train deep networks online with limited training samples, and inevitably suffer from overfitting. Transferring the hierarchical features learned for image classification tasks have been shown to be effective for numerous vision tasks, e.g., image segmentation [17], salient object detection [34]. More recent methods [26, 25, 11, 18] adopt deep convolution networks trained on a large scale image classification task [22] to improve tracking performance. The rich representation of transferred features from deep nets enables trackers to construct more robust, power appearance model over the traditional hand crafted feature based trackers.

**Spatio-temporal model based trackers:** Variations in the appearance of the object in tracking, such as variations in geometry/photometry, camera viewpoint, partial occlusion or out-of-view, pose a major challenge to object tracking. TLD [14] employs two experts to identify the false negatives and false positives to train a detector. The experts are independent, which ensures mutual compensation of their errors to alleviate the problem of drifting. A short and long term cognitive psychology principle is adopted in [12] to design a flexible representation that can adopt to changes in object appearance during tracking. A parameter-free Hedging algorithm is proposed in [2] for the problem of decision-theoretic online learning, especially for the applications when the number of actions is very large and optimally setting the parameter is not well understood. An improved Hedge algorithm considering historical performance is proposed in [21] to weight the decision from different CNN layers.

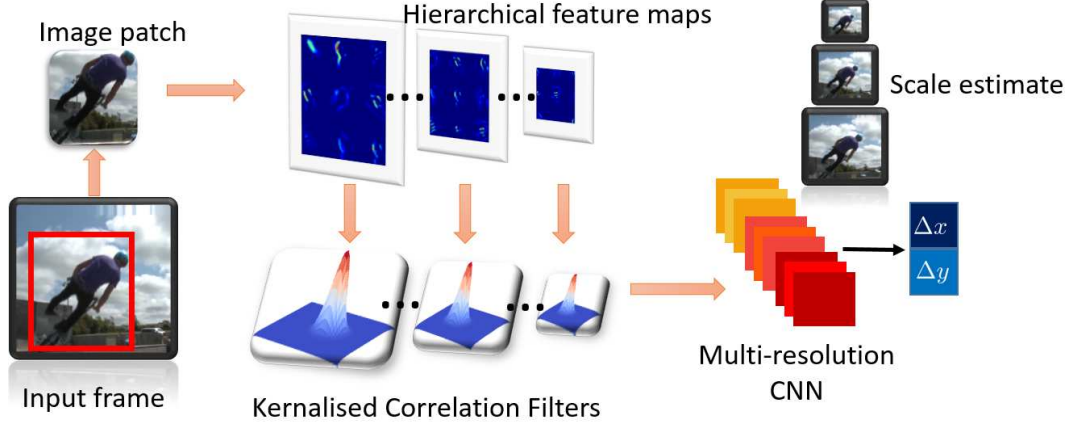


Figure 2: The proposed algorithm overview: we first extract hierarchical CNN features from the image patch of interest; then we project deep features into kernel space for correlation filters; the target translational movement is predicted via multi-resolution convnets and later we update the scale and appearance models using an adaptive learning rate scheme.

### 3. Proposed Algorithm

As shown in Fig. 2, the proposed approach consists of three steps: extracting hierarchical CNN features, projecting features into kernel space for correlation filters, predicting target translational movement via multi-resolution convnet. We first review the correlation filter as our building block. Then we present the technical details of the projection of deep feature to kernel space, the model of learning translational output using a multi-resolution convnets and an adaptive learning rate scheme for model update.

**Correlation Filter:** Correlation filters based trackers [1, 9, 3, 10] exploit the circulant structure of training and testing samples to greatly accelerate the training and testing process. Let  $\mathbf{X}^k \in \mathbb{R}^{D \times M \times N}$  denotes the feature sets where  $D$  denotes the number of feature maps;  $M, N$  denote the shape of feature maps and  $k$  denotes the  $k$ -th input feature map extracted from the  $k$ -th convolutional layers from a pretrained CNN;  $\mathbf{Y} \in \mathbb{R}^{M \times N}$  denotes a gaussian shaped label matrix which is subject to a 2D Gaussian distribution with zero mean and standard deviation proportional to the target size. The goal of training is to find a set of filters  $\mathbf{W}^k$  that minimises the squared error over sets of circulant translated samples  $\mathbf{X}^k$  and their regression targets  $\mathbf{Y}$ :

$$\mathbf{W}^k = \underset{\mathbf{W}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}^k \bullet \mathbf{W}^k\|^2 + \lambda \|\mathbf{W}^k\|^2 \quad (1)$$

where

$$\mathbf{X}^k \bullet \mathbf{W}^k = \sum_{d=1}^D \mathbf{X}_d^k \odot \mathbf{W}_d^k \quad (2)$$

with the symbol  $\odot$  denotes element-wise product. The minimizer has a closed-form:

$$\mathbf{W} = \frac{\mathbf{X}^T \mathbf{Y}}{\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}} \quad (3)$$

where  $\mathbf{I}$  is an identity matrix and  $\lambda$  is a regularisation parameter that controls overfitting. We drop the superscript  $k$  for notational simplicity. In general, a large system of linear equations must be solved to compute the solution, which can become prohibitive in a real-time setting. With training data being cyclic shifts patches, all operations can be done element-wise on their diagonal elements [10].

The filter can be modelled in the Fourier domain by:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X} \bullet \mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \quad (4)$$

where

$$\mathbf{X} \bullet \mathbf{W} = \sum_{d=1}^D \mathbf{X}_d \odot \mathbf{W}_d \quad (5)$$

The corresponding minimizer in the Fourier domain has the closed-form:

$$\mathbf{W} = \frac{\mathbf{X}^* \odot \mathbf{Y}}{(\mathbf{X}^* \odot \mathbf{X} + \lambda \mathbf{I})} \quad (6)$$

where  $*$  denotes the Hermitian transpose and since diagonal matrices are symmetric, taking the Hermitian transpose only left behind a complex-conjugate.

#### 3.1. Kernelisation of Deep transfered features

In Kernelized Correlation Filter (KCF) [10], the feature  $\mathbf{X}$  is mapped to a Hilbert space  $\phi(\mathbf{X})$ . By employing a kernel  $\kappa(\mathbf{X}, \mathbf{X}') = \langle \phi(\mathbf{X}), \phi(\mathbf{X}') \rangle$ , Eq. 4 becomes:

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{Y} - \langle \phi(\mathbf{X}), \mathbf{W} \rangle\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \quad (7)$$

The power of the kernel trick comes from the implicit use of a high-dimensional feature space  $\phi(\mathbf{X})$  without ever instantiating a vector in the space. Even though the regression

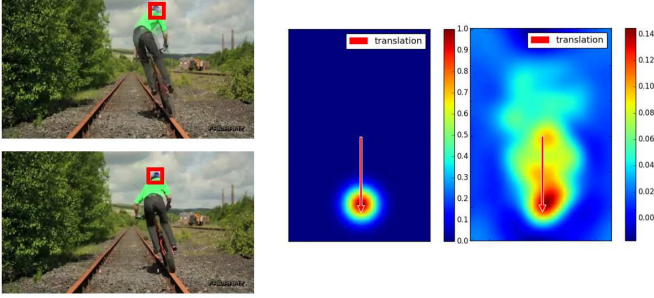


Figure 3: Response maps for one frame of the ‘‘Biker’’ sequence. Left: two consecutive frames; middle: ideal response map (2D gaussian); right: the actual response map. Arrows represent the translational vector. Due to various factors, *e.g.*, motion blur, target appearance change, scale change, boundary effect, the actual response map is not a univariate gaussian and the maximum response does not strictly correspond to the translational movement.

function’s complexity grows with the number of samples which is the major drawback of the kernel trick, assuming circulant data and the adopted kernel being shift invariant, the kernel correlation can be computed efficiently. In our experiment, an radial basis function (RBF) kernel, which satisfies the shift invariant property, is adopted:

$$\mathbf{k}^{\mathbf{X}\mathbf{X}'} = \exp\left(-\frac{1}{\sigma}(\|\mathbf{X}\|^2 + \|\mathbf{X}'\|^2 - 2\mathcal{F}^{-1}(\mathcal{X}^* \odot \mathcal{X}'))\right) \quad (8)$$

where  $\mathcal{F}^{-1}$  denotes the Inverse DFT and the full kernel correlation can be computed in only  $\mathcal{O}(n \log n)$  time.

Expressing the solution  $\mathcal{W}$  as a linear combination of the samples:  $\mathcal{W} = \sum_i \alpha_i \phi(\mathcal{X}_i)$  renders an alternative representation  $\alpha$  to be in the dual space, as opposed to the primal space  $\mathcal{W}$  (Representer Theorem [23]). The variables under optimisation are thus  $\alpha$ :

$$\alpha = (\mathbf{k}^{\mathbf{X}\mathbf{X}'} + \lambda \mathcal{I})^{-1} \mathbf{y} \quad (9)$$

For detection stage for translation estimation, given an image patch feature  $\mathbf{Z}^k \in \mathbb{R}^{D \times M \times N}$ , the response map is obtained by:

$$R(\mathbf{Z}^k) = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{k}^{\mathbf{X}\mathbf{Z}^k}) \odot \mathcal{F}(\alpha^k)) \quad (10)$$

The model for frame  $t$  is updated with learning rate  $\eta$  as:

$$\alpha_t^k = (1 - \eta)\alpha_{t-1}^k + \eta\alpha^k \quad (11)$$

### 3.2. Decoding Response map using Multi-Res CNN

Previous works estimate the next frame location via the maximum response point on the response map. We argue that albeit the maximum response point carries physical meaning of the translational information between two

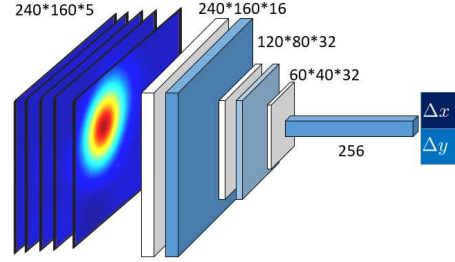


Figure 4: A multi-resolution convnet is deployed to decode the hierarchical response maps.

consecutive frames, the precondition is that of static target appearance, *i.e.*, the maximum response map corresponds to the cyclic shifts of appearance unchanged target. Due to subtle changes of target appearance, *e.g.*, rotation, scale, cyclic boundary effect or motion blur of two consecutive frames, the accuracy of finding the translation using maximum response could be compromised (*c.f.* Fig. 3).

Moreover, when using deep features from the pre-trained convnets for the input representation  $\mathbf{X}$ , the hierarchies of feature maps capture different semantic information (*c.f.* Fig. 1): features from deep layers capture rich category level semantic information, which is useful for object classification, but they are not the optimal representation for visual tracking because spatial details captured by earlier layers are also important for accurately localising the targets. On the other hand, as the features in the earlier layers capture low level visual characteristics and are more class-generic rather than discriminative as ones in the later layers, methods based on features from earlier layers are likely to fail in challenging scenarios when the target size is small and high level object is the target of tracking.

We interpret the response maps from the hierarchies of convolutional layers as a nonlinear counterpart of an image pyramid representation and employ a learning framework for mapping the hierarchical outputs to target translation. A multi-resolution neural network (*c.f.* Fig. 4) is deployed to exploit the implicit translational information from hierarchies response maps:

$$\Delta x, \Delta y \leftarrow \text{convnets}(R(\mathbf{Z}^k)) \quad (12)$$

The loss of the convnets is set to be the root mean square (RMS) of the normalised translational movement  $(\Delta x, \Delta y)$  and the predicted movement  $(\Delta x', \Delta y')$ :

$$L_{pos} = \sqrt{\frac{1}{2}(\|\Delta x - \Delta x'\|_2^2 + \|\Delta y - \Delta y'\|_2^2)} \quad (13)$$

The proposed formulation enables efficient integration of multi-resolution response maps to decode translational information implicitly and demonstrates its competitiveness over the corresponding counterpart baselines that using the

mean of maximum responses as the indicator for target movement.

### 3.3. Adaptive learning rate

Traditional correlation based filter updates the model by a fixed parameter  $\eta$ . For generic object tracking, however, there are two crucial factors for model update using deep features: (1) the appearance of object of interest usually changes at irregular pace (sometimes gently and sometimes vehemently). This means that the scale of the learning rate should reflect the appearance change of the target; (2) depending on whether the object of interest being a low level visual cues or high level object entity, the hierarchies of response maps render correspondingly different maximum response values.

Let the ultimate target position is predicted  $(x_p, y_p)$  at time  $t$ , each layer  $k$  will incur a loss from its response map  $R_t^k$ :

$$l_t^k = \max(R_t^k) - R_t^k(x_p, y_p) \quad (14)$$

We then measure the stability of layer  $k$  at time  $t$ :

$$s_t^k = \frac{|l_t^k - \mu_t^k|}{\sigma_t^k} \quad (15)$$

where  $\mu_t^k, \sigma_t^k$  are the mean and variance for the loss  $l_t^k$  during time period  $\Delta t$ .

A larger  $s_t^k$  indicates the layer is less correlated with the object at frame  $t$  hence the model update  $\eta$  should be decreased as well. Therefore, we propose an adaptive learning rate for different layers of model update that is linear to the model stability as:

$$\eta^k = s_t^k \times \eta \quad (16)$$

Algorithm 1 summarises the main steps of the proposed approach for visual tracking.

## 4. Experiments

We validate our proposed KMC framework by evaluating two genres of experiments: one compares with correlation filter based baseline trackers and one compares with several state-of-the-art trackers. We perform comprehensive experiments on three datasets: OTB-2013 [31], OTB-2015 [32] and UAV123 [19].

**Implementation details:** For feature extraction, we crop the image patch with 2.2 padding size and resize the image patch to  $240 * 160$  because the average target size is of ratio  $\frac{3}{2}$ . Feature bandwidth  $\sigma$  in Eq. 8 is 0.2. The learning rate  $\eta$  for Eq. 16 is 0.0025 which is only one fifth of ones chosen in [21] because of the model is in kernel space and tends to be more stable. After the forward propagation, we use the VGG-Net with 19 layers and five outputs before the max pooling layers (*i.e.*, ‘*block1\_conv2*’,

---

### Algorithm 1: Kernalised Multi-resolution Convnet

---

**Input:** target position  $(x_1, y_1)$  and size  $(w_1, h_1)$ , first tracking frame;

**Output:** predicted target positions  $(x_t, y_t)$  and sizes  $(w_t, h_t)$  in the following frames.

```

1 for  $t = 1, 2, \dots$  do
2   Crop target images with padding;
3   Extract deep features from layers before max
   pooling layers in VGG-Net19 to obtain 5
   hierarchical representations  $\mathbf{X}$ ;
4   Project deep features into kernel space 3.1;
5   if  $t = 1$  then
6     Obtain deep kernel model  $\alpha$  (Eq.9);
7   else
8     Update target position prediction by the
     trained multi-resolution convnet 3.2;
9     Update model scale using DSST [4];
10    Update kernel model using adaptive learning
     rate 3.3;

```

---

‘*block2\_conv2*’, ‘*block3\_conv4*’, ‘*block5\_conv4*’). Note that instead of heuristically choosing layers as in [21, 7], we have included all layer outputs before the max pooling layers. This chosen methodology makes the approach more generic and less dataset dependant.

The convnet in Sec. 3.2 is trained on OTB-2015 dataset. Note that albeit trained on the OTB-2105 dataset, the convnet during detection will not be able to see the exactly same response maps twice unless exact location update is achieved. Hence it requires the convnet to be able to generalise to unseen hierarchical response maps and predict reasonable translational output. Furthermore, we verify the generalisation of the KMC over HDT [21] tracker on the unseen UAV123 dataset.

For scale estimation, we use a scale pyramid representation as in [4] with 11 scales and a relative scale factor 1.02. The updates consistently helps more accurate update of target models while maintaining the computational cost low.

**Evaluation Methodology:** Following the evaluation strategy of [31], all trackers are compared using two measures: precision and success. Precision is measured as the distance between the centres of the ground truth bounding box and the corresponding tracker generated bounding box. The precision plot shows the percentage of tracker bounding boxed within a given threshold distance in pixels of the ground truth. To rank the trackers, the conventional threshold of 20 pixels (*P20*) is adopted. Success is measured as the intersection over union of pixels. The success plot shows the percentage of tracker bounding boxes whose overlap score is larger than a given threshold and the track-



Method \ Dataset	OTB-2013		OTB-2015	
	P20	AUC	P20	AUC
DeepDCF	58.0	38.4	64.3	44.0
<b>DeepKCF</b>	66.2 (8.2 ↑)	46.1 (7.7 ↑)	71.4 (7.1 ↑)	49.1 (5.1 ↑)
MaxRes	57.5	38.8	65.8	45.9
<b>Multi-Resolution CNN</b>	70.6 (13.1 ↑)	46.7 (7.9 ↑)	77.8 (12.0 ↑)	53.6 (7.7 ↑)
Fixed Learning Rate	74.0 (3.4 ↑)	51.8 (5.1 ↑)	79.8 (2.0 ↑)	57.0 (3.4 ↑)
<b>Adaptive Learning Rate</b>	74.2 (0.2 ↑)	52.1 (0.3 ↑)	79.9 (0.1 ↑)	57.4 (0.4 ↑)

Table 1: Baseline comparisons with arrows indicating performance changes comparing with the upper row parallel experiment. First two rows investigate the impact of projecting deep features into a kernel space vs. using linear deep features. Next two rows compare the traditional using max location vs. the proposed multi-resolution convnets paradigm for decoding target translational movement. Last two rows illustrate the benefit of adaptive learning rate through time (with scale update).

ers are ranked according to the Area Under Curve (AUC) criteria. All sequences are evaluated using One-Pass Evaluation (OPE) as in [31].

#### 4.1. Baseline comparison

**Benefit of Kernalsing deep features:** We first evaluate the impact of projecting deep features into a kernel space (*DeepKCF*) vs. using the traditional linear projection (*DeepDCF*) in first two rows of Tab. 1. The deep features used are the layer from ‘*block3\_conv4*’. It can be seen that projecting deep feature into kernel space consistently yields a noticeable increase in both performance measures. The computation cost for kernel projection is  $\mathcal{O}(n \log n)$  and for linear kernel is  $\mathcal{O}(n)$ . Therefore, given marginal increase in computational cost, we conclude that for constructing appearance model, kernel projection is the preferred methodology versus the traditional linear projection.

**Benefit of multi-resolution CNN:** We then evaluate the impact of decoding response map using a multi-resolution CNN (*Multi-Resolution CNN*) vs. directly using the max location information (*MaxRes*) for decoding target location from hierarchical response maps in Tab. 1. *MaxRes* use the maximum position from the mean of hierarchical response maps with later response maps being resized to the largest first layer response map of size  $240 * 160$ . The performance of *MaxRes* is even worse than using single layer, signifying the need to intelligently combine multiply layer output. The learnt convnets consistently performs better than that of *MaxRes* for a large margin. Moreover, the learning paradigm opens the door for the integration of recurrent network into the pipeline.

**Benefit of adaptive learning rate through time:** We also investigate the scheme of adjusting the learning rate adaptively vs. fixing the learning rate through time in last two rows of Tab. 1. Consistent improvements across two metrics

are observed. The margin of improvement, however, is less prominent. We reason that in part, it is due to almost saturate performance; on the other hand, an end-to-end learning scheme for model update could be more preferable and should be further investigated.

#### 4.2. Comparisons with the-state-of-the-art trackers

We validate our KMC tracker in a comprehensive comparison with 10 state-of-the-art trackers: HDT [21], SRDCF [6], MEEM [33], MUSTER [12], SAMF [16], DSST [4], KCF [10], Struck [8], TLD [13], DCF [9].

**OTB-2015 Dataset:** OTB-2015 dataset contains 100 video sequences and is the superset of OTB-2013 [31] dataset which contains the original 50 video sequences. The Results on the OBT dataset is shown in the top row of Fig. 5. It can be seen that our tracker performs on par with a range of state-of-the-art trackers across two evaluation metrics. The most similar tracker is HDT where both correlation filter and deep features are adopted. However, HDT has heuristically chosen a set of higher level deep feature layers whereas in our approach, all layers before max pooling are universally chosen. The SRDCF tracker is better than our tracker in the success metric due to heavy regularisation for negative training examples in SRDCF tracker. Hence it enables larger search area during tracking. Furthermore, because SRDCF tracker is able to search for larger area during detection ( $4^2$  vs. our  $3.2^2$  the target size), it indirectly manages the training set to handle the scenarios of occlusion and drift, of which the proposed KMC has been mostly penalised.

**UAV123 Dataset:** Visual tracking on unmanned aerial vehicles (UAVs) is a very promising application, since the camera can follow the target based on visual feedback and actively change its orientation and position to optimise the tracking performance. This UAV123 dataset contains a total

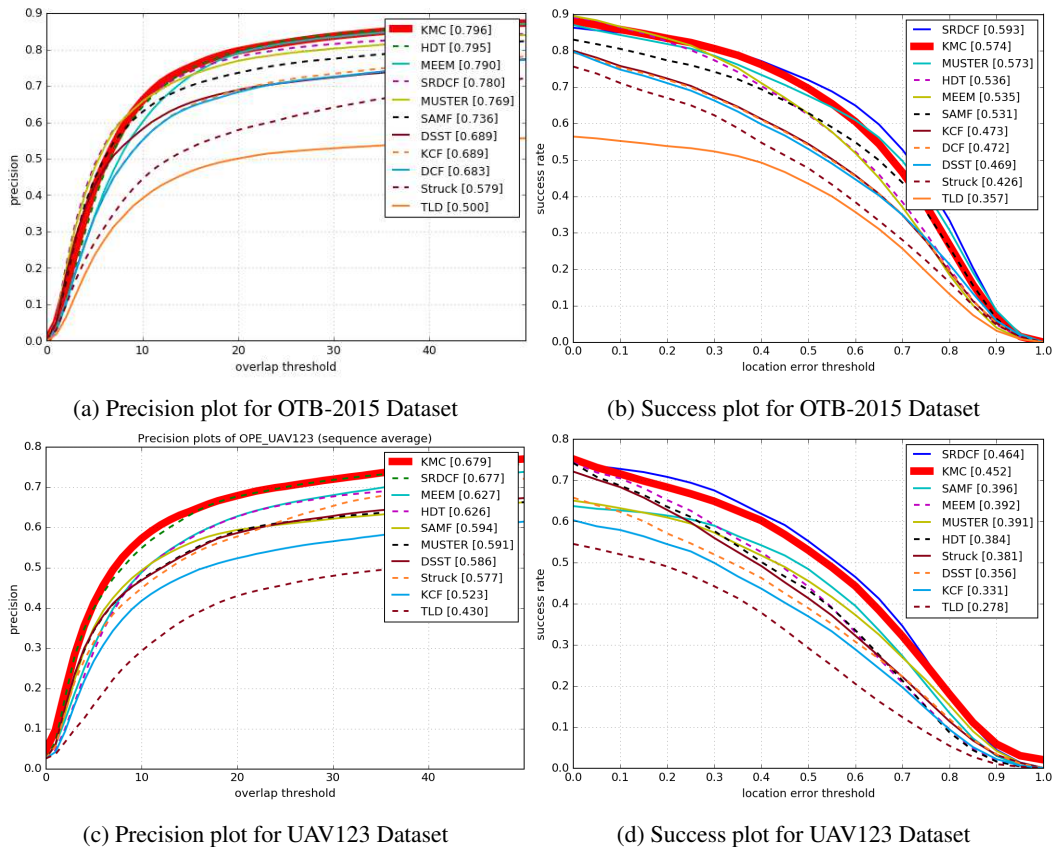


Figure 5: Precision plot and success plot of the state-of-the-art trackers for OTB-2015 and UAV123 dataset.

of 123 video sequences and more than 110K frames. The major difference between this UAV123 dataset and other popular tracking datasets is the effect of camera viewpoint change arising from UAV motion, the variation in bounding box size and aspect ratio with respect to the initial frame, and longer tracking sequences on average due to the availability of mounted camera moving with the target.

Results are shown in the bottom row of Fig. 5. It can be seen that due to heuristically chosen layers tuned to the OBT dataset in HDT, its performance suffers on this UAV123 dataset. It worths noting that our trained multi-resolution convnet has never seen any image in this dataset. Moreover, no pesky threshold, hyperparameters nor weights are altered when applying the trained KMC network on the unseen UAV123 dataset, proving its ability to generalise across various tracking scenarios.

## 5. Conclusions

In this paper, we propose a novel kernelised multi-resolution convnet tracking algorithm that utilises the intermediate response maps from the kernelised correlation filter outputs. The multi-resolution convnet learns the implicit translational output accurately and later an adaptive

learning scheme is adopted for model update. The learning paradigm is able to generalise across various datasets without change of hyperparameters. Moreover, it opens the door on the end-to-end temporal deep learning. Future works include better regularisation method as in [6] for negative instance mining, multi-layer fusion [24] and incorporating recurrent nets on top to model temporal dynamics.

## Details of the Code

The python based code can be found at:  
[https://github.com/stevenwudi/KMC\\_cvprw\\_2017](https://github.com/stevenwudi/KMC_cvprw_2017)

## Acknowledgment

This project has received funding from National Natural Science Foundation of China (NSFC) (61401287); Natural Science Foundation of Shenzhen (JCYJ20160307154003475, JCYJ2016050617265125).

## References

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

- [2] K. Chaudhuri, Y. Freund, and D. J. Hsu. A parameter-free hedging algorithm. In *Advances in neural information processing systems*, 2009.
- [3] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*, 2014.
- [4] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Discriminative scale space tracking. In *Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [5] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. *International Conference on Computer Vision, workshop*, 2015.
- [6] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. *International Conference on Computer Vision*, 2015.
- [7] M. Danelljan, K. F. S. Robinson, Andreas, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. *European Conference on Computer Vision*, 2016.
- [8] S. Hare, A. Saffari, and P. Torr. Struck: Structured output tracking with kernels. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.
- [9] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, 2012.
- [10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
- [11] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. *International Conference on Machine Learning*, 2015.
- [12] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [13] Z. Kalal, J. Matas, and K. Mikolajczyk. P-n learning: Bootstrapping binary classifiers by structural constraints. In *Conference on Computer Vision and Pattern Recognition*, 2010.
- [14] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 2012.
- [15] H. Li, Y. Li, and F. Porikli. Robust online visual tracking with a single convolutional neural network. In *Asian Conference on Computer Vision*, 2014.
- [16] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. *European Conference on Computer Vision, workshop*, 2014.
- [17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *Conference on Computer Vision and Pattern Recognition*, 2015.
- [18] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *IEEE International Conference on Computer Vision*, 2015.
- [19] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, 2016.
- [20] P. Ondruška and I. Posner. Deep tracking: Seeing beyond seeing using recurrent neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [21] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, and J. L. M.-H. Yang. Hedged deep tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.
- [23] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [24] L. Shao, D. Wu, and X. Li. Learning deep and wide: A spectral method for learning deep networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.
- [25] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- [26] L. Wang, W. Ouyang, X. Wang, and H. Lu. Stct: Sequentially training convolutional networks for visual tracking. In *IEEE International Conference on Computer Vision*, 2016.
- [27] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *Advances in Neural Information Processing Systems* 26. 2013.
- [28] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *Advances in neural information processing systems*, 2013.
- [29] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2016.
- [30] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [31] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *IEEE conference on computer vision and pattern recognition*, 2013.
- [32] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
- [33] M.-S.-S. S. Zhang, J. Meem: robust tracking via multiple experts using entropy minimization. In *European Conference on Computer Vision*, 2014.
- [34] W. Zou and N. Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.