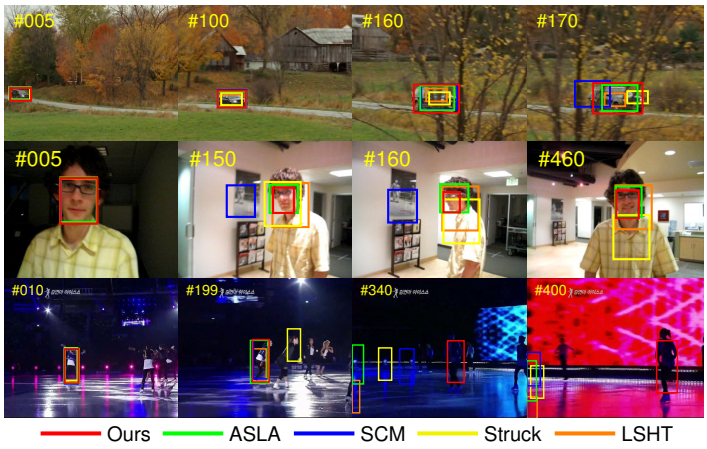
Accurate Scale Estimation for Robust Visual Tracking

Martin Danelljan, Gustav Häger,
Fahad Shahbaz Khan, Michael Felsberg
martin.danelljan@liu.se, hager.gustav@gmail.com,
fahad.khan@liu.se, michael.felsberg@liu.se

Computer Vision Laboratory
Department of Electrical Engineering
Linköping University
Linköping, Sweden

Robust scale estimation is a challenging problem in visual object tracking. Most existing methods fail to handle large scale variations in complex image sequences. This paper presents a novel approach for robust scale estimation in a tracking-by-detection framework. The proposed approach works by learning discriminative correlation filters based on a scale pyramid representation. We learn separate filters for translation and scale estimation, and show that this improves the performance compared to an exhaustive scale search while operating at real-time. Our scale estimation approach is generic as it can be incorporated into any tracking method with no inherent scale estimation.



Discriminative Correlation Filters. Our tracking approach is based on the discriminative correlation filters employed in the MOSSE tracker [1]. Similarly to [2], these filters are extended to multi-dimensional features for visual tracking. We use HOG features for the translation filter and concatenate it with image intensity features. In general, we consider a d-dimensional feature map representation of an image. Let f be a rectangular patch of the target, extracted from this feature map. We denote feature dimension number $l \in \{1,\dots,d\}$ of f by f^l. The objective is to find an optimal correlation filter h, consisting of one filter h^l per feature dimension. This is achieved by minimizing the cost function:

$$\varepsilon = \left\| \sum_{l=1}^{d} h^l \star f^l - g \right\|^2 + \lambda \sum_{l=1}^{d} \|h^l\|^2. \quad (1)$$

Here, g is the desired correlation output associated with the training example f and $\lambda \geq 0$ is a regularization parameter. The solution to (1) is:

$$H^l = \frac{\overline{G}F^l}{\sum_{k=1}^{d} \overline{F^k}F^k + \lambda}. \quad (2)$$

Capital letters denote the discrete Fourier transforms (DFTs) of the corresponding functions. We update the numerator A_t^l and denominator B_t of the correlation filter H_t^l in (2) separately using a learning rate η:

$$A_t^l = (1-\eta)A_{t-1}^l + \eta \overline{G}_t F_t^l \quad \text{and} \quad B_t = (1-\eta)B_{t-1} + \eta \sum_{k=1}^{d} \overline{F_t^k} F_t^k. \quad (3)$$

The correlation scores y at a patch z in the next frame are computed using (4). The new target state is found by maximizing the score y.

$$y = \mathscr{F}^{-1}\left\{ \frac{\sum_{l=1}^{d} \overline{A_t^l} Z^l}{B_t + \lambda} \right\}. \quad (4)$$

Our Scale Estimation Approach. Ideally, an accurate scale estimation approach should be robust while computationally efficient. To achieve this, we propose a fast scale estimation approach by learning separate filters for translation and scale. This helps by restricting the search area to smaller parts of the scale space. In addition, we gain the freedom of selecting the feature representation for each filter independently.

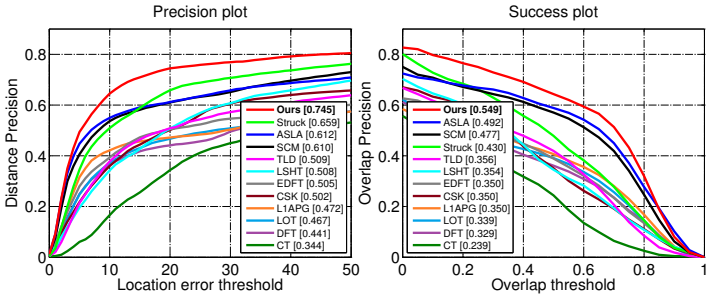


Figure 1: Precision and success plots illustrating the average distance and overlap precision respectively over all the 28 sequences. The average distance precision at 20 pixels for each method is reported in the legend of the precision plot. The legend of the success plot contains the *area-under-the-curve* (AUC) score for each tracker.

Method	median OP	median DP	median CLE	median FPS
Baseline (no scale)	37.8	74.5	15.9	**44.1**
Exhaustive Scale Search (this paper)	52.2	87.6	11.8	0.96
Fast Scale Search (this paper)	**75.5**	**93.3**	**10.9**	24.0

Table 1: Comparison of our fast scale estimation method with the baseline tracker and our exhaustive scale-space tracker.

We augment the baseline method by learning a separate 1-dimensional correlation filter to estimate the target scale in an image. The training example f for updating the scale filter is computed by extracting features using variable patch sizes centred around the target. Let $P \times R$ denote the target size in the current frame and S be the size of the scale filter. For each $n \in \left\{ \left\lfloor -\frac{S-1}{2} \right\rfloor, \dots, \left\lfloor \frac{S-1}{2} \right\rfloor \right\}$, we extract an image patch J_n of size $a^n P \times a^n R$ centred around the target. Here, a denotes the scale factor between feature layers. The value $f(n)$ of the training example f at scale level n is set to a HOG-based d-dimensional feature descriptor of J_n. Eq. 3 is then used to update the scale filter h_{scale} with the new sample f.

In visual tracking scenarios, the scale difference between two frames is typically smaller compared to the translation. Therefore, we first apply the translation filter h_{trans} given a new frame. Afterwards, the scale filter h_{scale} is applied at the new target location. An example z is extracted from this location using the same procedure as for f. By maximizing the correlation output (4) between h_{scale} and z, we obtain the scale difference.

Evaluation. We employ all the 28 sequences annotated with the scale variation attribute in the recent evaluation of tracking methods [3]. The sequences also pose challenging problems such as illumination variation, motion blur, background clutter and occlusion. The baseline HOG based tracker with no scale estimation capability is compared with our exhaustive scale space tracker and the fast scale estimation method in table 1.

We additionally compare our approach with 11 state-of-the-art trackers. Figure 1 contains the precision and success plots illustrating the *mean* distance and overlap precision over all the 28 sequences. In both precision and success plots, our approach significantly outperforms the compared methods. In summary, the precision plot demonstrates that our approach is superior in robustness compared to existing trackers. Similarly, the success plot shows that our method estimates the target scale more accurately on the benchmark sequences.

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Yui M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.

[2] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *CoRR*, abs/1404.7584, 2014.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013.