Long-term Correlation Tracking

Chao Ma[1,2], Xiaokang Yang[1], Chongyang Zhang[1], and Ming-Hsuan Yang[2]
[1]Shanghai Jiao Tong University. [2]University of California at Merced.

In this paper, we address the problem of long-term visual tracking where the target objects undergo significant appearance variation due to deformation, abrupt motion, heavy occlusion and out-of-view. Our approach builds on two major observations based on prior work. First, it is important to model the temporal relationship of appearance consisting of a target object and its context as there is little change between two consecutive frames. Second, it is critical to enhance the detection module of a long-term tracker to (i) estimate the scale change and (ii) re-detect the object in case of tracking failure when the long-term occlusion or out-of-view occur.

We decompose the task of tracking into translation and scale estimation of objects. We show that the correlation between temporal context considerably improves the accuracy and reliability for translation estimation, and it is effective to learn the discriminative correlation filters from the most confident frames to estimate the scale change. A typical tracker [2, 3, 9] based on correlation filters models the appearance of a target object using a filter $\mathbf{w}$ trained on an image patch $\mathbf{x}$ of $M \times N$ pixels, where all the circular shifts of $\mathbf{x}_{m,n}$, $(m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}$, are generated as training samples with Gaussian function label $y(m,n)$, i.e.,

$$\mathbf{w} = \underset{\mathbf{w}}{\mathrm{argmin}} \sum_{m,n} |\phi(\mathbf{x}_{m,n}) \cdot \mathbf{w} - y(m,n)|^2 + \lambda |\mathbf{w}|^2, \tag{1}$$

where ϕ denotes the mapping to a kernel space and λ is a regularization parameter ($\lambda \geq 0$). Since the label $y(m,n)$ is not binary, the learned filter $\mathbf{w}$ contains the coefficients of a Gaussian ridge regression [6] rather than a binary classifier. Using the fast Fourier transformation (FFT) to compute the correlation, this objective function is minimized as $\mathbf{w} = \sum_{m,n} \mathbf{a}(m,n)\phi(\mathbf{x}_{m,n})$, and the coefficient $\mathbf{a}$ is defined by

$$A = \mathcal{F}(\mathbf{a}) = \frac{\mathcal{F}(\mathbf{y})}{\mathcal{F}(\phi(\mathbf{x}_{m,n}) \cdot \phi(\mathbf{x})) + \lambda}. \tag{2}$$

where $\mathcal{F}$ denotes the discrete Fourier operator and $\mathbf{y} = \{y(m,n) | (m,n) \in \{0,1,\ldots,M-1\} \times \{0,1,\ldots,N-1\}\}$. The tracking task is carried out on an image patch $\mathbf{z}$ in the new frame with the search window size $M \times N$ by computing the response map as $\hat{\mathbf{y}} = \mathcal{F}^{-1}\left(A \odot \mathcal{F}(\phi(\mathbf{z}) \cdot \phi(\hat{\mathbf{x}}))\right)$, where $\hat{\mathbf{x}}$ denotes the learned target appearance model and $\odot$ is the Hadamard product. Therefore, the new position of target is detected by searching for the location of the maximal value of $\hat{\mathbf{y}}$.

Differently from prior work, we train two regression models based on correlation filters from one single frame. As shown in Fig. 1, the temporal context model R_c takes both the target and surrounding context into account, since this information remains temporally stable and useful to discriminate the target from the background in the case of occlusion [9]. It is important for the regression model R_c to be adaptive to estimate the translation when the target undergoes occlusion, deformation, and abrupt motion. The R_c model is thus updated with a learning rate α frame by frame as

$$\begin{aligned} \hat{\mathbf{x}}^t &= (1-\alpha)\hat{\mathbf{x}}^{t-1} + \alpha \mathbf{x}^t, \\ \hat{A}^t &= (1-\alpha)\hat{A}^{t-1} + \alpha A^t, \end{aligned} \tag{3}$$

where t is the index of the current frame.

We learn another discriminative regression model R_t from the most reliable tracked targets. Specifically, we use the maximal value of $\hat{\mathbf{y}}$ to determine the confidence of tracking results. To maintain the model stability, we use a pre-defined threshold $\mathcal{T}_a$ and only update R_t using (3) if $\max(\hat{\mathbf{y}}) \geq \mathcal{T}_a$. Note that there are no cosine spatial weights for model R_t in the feature space (See Fig. 1). During tracking, we construct a target pyramid around the estimated translation location for scale estimation. Let $P \times Q$ be the target size in a test frame and N indicate the number of scales $S = \{a^n | n = \lfloor -\frac{N-1}{2} \rfloor, \lfloor -\frac{N-3}{2} \rfloor, \ldots, \lfloor \frac{N-1}{2} \rfloor\}$. For each $s \in S$, we extract an image patch J_s of size $sP \times sQ$ centered around the estimated location. Unlike [1], we propose to uniformly resize all patches with size $P \times Q$ again and use HOG features to construct the scale feature pyramid. Let $\hat{\mathbf{y}}_s$ denote the correlation response map of the target regressor R_t to J_s, the optimal scale $\hat{s}$ of target is $\hat{s} = \mathrm{argmax}_s \left(\max(\hat{\mathbf{y}}_1), \max(\hat{\mathbf{y}}_2), \ldots, \max(\hat{\mathbf{y}}_S)\right)$. Accordingly, the regression model R_t is updated by (3) if $\max(\hat{\mathbf{y}}_{\hat{s}}) \geq \mathcal{T}_a$.

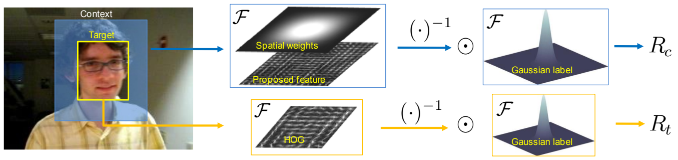

Figure 1: Two regression models learned from a single frame. The model R_c exploits the temporal correlation of target and surrounding context while R_t only models target appearance. To train the model R_c, a layer of spatial weights are added on the feature space. Here $\mathcal{F}$ denotes the discrete Fourier operator and $\odot$ is the Hadamard product.

It is clear that a robust long-term tracking algorithm requires a re-detection module in the case of tracking failure. Different from previous trackers [4, 8], where re-detection is carried out on each frame, we use a threshold $\mathcal{T}_r$ to activate the detector if $\max(\hat{\mathbf{y}}_{\hat{s}}) < \mathcal{T}_r$. For computational efficiency, we do not use the regression model R_t as a detector and instead use the online random fern classifier [5]. As the detector is applied to the entire frame with sliding windows when $\max(\hat{\mathbf{y}}_{\hat{s}}) < \mathcal{T}_r$, we train an online random ferns detector with a conservative update scheme. Let c_i, $i \in \{0,1\}$ be the indicator of class labels and let f_j, $j \in \{1,2,\ldots,N\}$ be the set of binary features, which are grouped into small sets as ferns. The joint distribution for features in each fern is $P(f_1, f_2, \ldots, f_N | C = c_i) = \prod_{k=1}^{M} P(F_k | C = c_i)$, where $F_k = \{f_\sigma(k,0), f_\sigma(k,2), \ldots, f_\sigma(k,N)\}$ represents the k-th fern, and $\sigma(k,n)$ is a random permutation function with range from 1 to N. For each fern F_k, its conditional probability can be written as $P(F_k | C = c_i) = \frac{N_{k,c_i}}{N_k}$, where N_{k,c_i} is the number of training samples of class c_i that belongs to the k-th fern and N_k is the total number of training samples that fell into the leaf-node corresponding to the k-th fern. From the Bayesian perspective, the optimal class $\hat{c}_i$ is detected as $\hat{c}_i = \mathrm{argmax}_{c_i} \prod_{k=1}^{M} P(F_k | C = c_i)$ [7].

Extensive experimental results on large-scale benchmark datasets show that the proposed algorithm performs favorably against state-of-the-art methods in terms of efficiency, accuracy, and robustness.

[1] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014.

[2] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, 2014.

[3] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, In Preprint, 2015.

[4] Y. Hua, K. Alahari, and C. Schmid. Occlusion and motion reasoning for long-term tracking. In *ECCV*, 2014.

[5] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *TPAMI*, 34(7):1409–1422, 2012.

[6] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[7] M. Özuysal, P. Fua, and V. Lepetit. Fast keypoint recognition in ten lines of code. In *CVPR*, 2007.

[8] J. S. Supancic and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013.

[9] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.