Enable Scale and Aspect Ratio Adaptability in Visual Tracking with Detection Proposals

Dafei Huang[1][2]
huangdafei1012@163.com
Lei Luo[†][1]
l.luo@nudt.edu.cn
Mei Wen[1][2]
meiwen@nudt.edu.cn
Zhaoyun Chen[1][2]
chenzhaoyun09@163.com
Chunyuan Zhang[1][2]
cyzhang@nudt.edu.cn

[1] College of Computer
National University of Defense Technology
Changsha, China

[2] National Key Laboratory of Parallel and Distributed Processing
National University of Defense Technology
Changsha, China

Among increasingly complicated trackers in visual tracking area, recently proposed correlation filter based trackers have achieved appealing performance despite their great simplicity and superior speed. However, the filter input is a bounding box of fixed size, so they are not born with the adaptability to target's scale and aspect ratio changes. Although scale-adaptive variants have been proposed, they are not flexible enough due to pre-defined scale sampling manners. Moreover, to the best of our knowledge, no correlation filter variant has been proposed to handle aspect ratio variation. To tackle this problem, this paper integrates the class-agnostic detection proposal method, which is widely adopted in object detection area, into a correlation filter tracker, and presents KCFDP tracker.

The correlation filter part of KCFDP is based on KCF[2] with some modifications. We extend the HOG feature in KCF to a combination of HOG, intensity, and color naming by simply concatenating the three features, resulting in 42 feature channels. The model updating scheme in KCF, which is simple linear interpolation, is substituted with a more robust scheme presented in [1]. EdgeBoxes[4] is adopted to generate flexible detection proposals and enable the scale and aspect ratio adaptability of our tracker. It traverses the whole image in a sliding window manner, and scores every sampled bounding box according to the number of contours that are wholly enclosed. To accelerate EdgeBoxes and produce less unnecessary proposals, we set the minimum proposal area and aspect ratio range dynamically in sliding window sampling according to the current target size.

In the tracking pipeline, KCF is firstly performed to estimate the preliminary target location $\mathbf{l}^d$. Within a patch $\mathbf{z}^d$ extracted from current frame, KCF locates the target center according to the location of the maximum element in $\mathbf{f}$:

$$\hat{\mathbf{f}}(\mathbf{z}^d) = \hat{\mathbf{k}}^{\bar{\mathbf{x}}\mathbf{z}^d} \cdot \hat{\alpha}, \qquad (1)$$

where ˆ denotes the discrete Fourier transform (DFT). α and $\bar{\mathbf{x}}$ are the coefficient matrix and current target appearance respectively. $\mathbf{k}$ represents the kernel correlation operation defined as:

$$\mathbf{k}^{\mathbf{x}'\mathbf{x}''} = \exp\left(-\frac{1}{\sigma^2}\left(\left\|\mathbf{x}'\right\|^2 + \left\|\mathbf{x}''\right\|^2 - 2\mathcal{F}^{-1}\left(\sum_c \hat{\mathbf{x}}'^{*}_{c} \cdot \hat{\mathbf{x}}''_{c}\right)\right)\right), \qquad (2)$$

where σ is the bandwidth of Gaussian kernel, $\mathcal{F}^{-1}$ denotes the inverse DFT and $*$ refers to complex conjugation. The subscript c here means the cth channel of feature.

Then EdgeBoxes is performed on a patch $\mathbf{z}^p$ centered at the preliminary location $\mathbf{l}^d$, and of size slightly larger than the current estimated target size. We only take the top 200 proposals and further filter them with proposal rejection: if the intersection over union (IoU) between a proposal and the current detected target is higher than 0.9 or lower than 0.6, the proposal is rejected. We then evaluate each accepted proposal $\mathbf{p}$ using:

$$f(\mathbf{p}) = sum(\mathbf{k}^{\bar{\mathbf{x}}\mathbf{p}} \cdot \alpha), \qquad (3)$$

where $sum()$ means the summation of all the elements in a matrix. The proposal with maximum f, whose location and size are denoted as $\mathbf{l}^p$ and (w^p, h^p), is the most promising candidate, and thus adopted to update the target location $\mathbf{l}$ and size (w, h) with a damping factor γ:

$$\mathbf{l} = \mathbf{l}^d + \gamma(\mathbf{l}^p - \mathbf{l}^d); \quad (w,h) = (w,h) + \gamma((w^p,h^p) - (w,h)). \qquad (4)$$

The dataset adopted to evaluate our KCFDP is from [3], which consists of 50 sequences (including 51 tracking targets) with many challenging attributes. 29 trackers from [3] and five additional correlation filter trackers are included in comparison. Three experiments are conducted on a 28-sequence subset with significant scale variation, a 14-sequence subset with dramatic aspect ratio variation, and the whole dataset respectively. Evaluation results are provided in two kinds of plots: Precision Plot that indicates the ratio of frames with center location error (CLE) below a certain threshold, where trackers are ranked using a threshold of 20 pixels, and Success Plot that indicates the percentage of frames with IoU larger than a threshold comparing to ground truth, where trackers are ranked using the area under curve (AUC). KCFDP reports the highest accuracy in all the experiments as shown in Fig.1, while running efficiently at an average speed of 20.8 FPS.

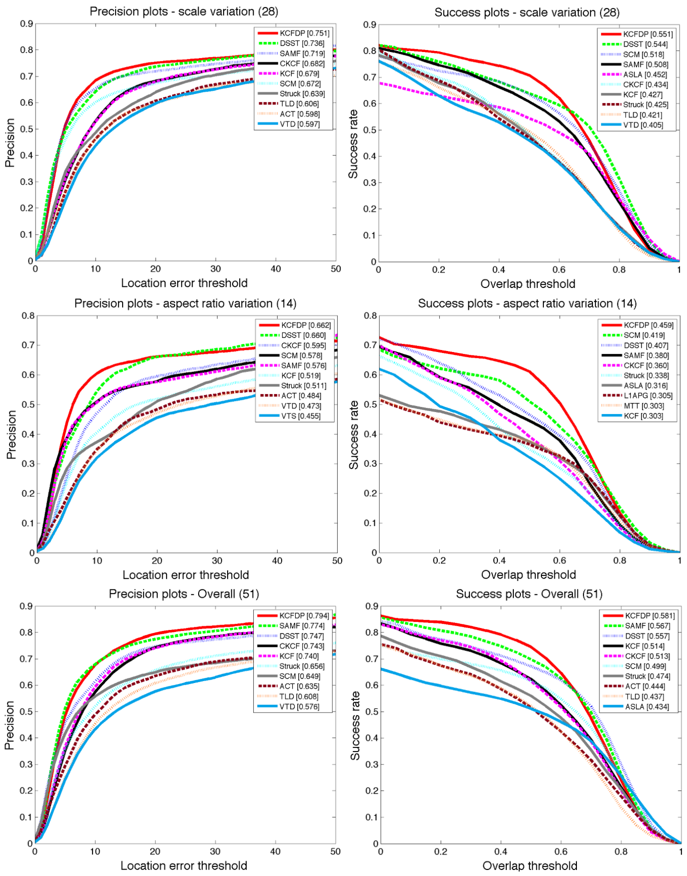

Figure 1: Evaluation results for scale adaptability (upper), aspect ratio adaptability (middle), and on the whole dataset (lower).

[1] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, pages 1090–1097, 2014.

[2] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015. doi: 10.1109/TPAMI.2014.2345390.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013.

[4] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405, 2014.