LEARNING A SCALE-AND-ROTATION CORRELATION FILTER FOR ROBUST VISUAL TRACKING

Yan Li, Guizhong Liu, Member IEEE

School of Electronic and Information Engineering
Xi'an Jiaotong University, Xi'an, China 710049

ABSTRACT

Robust scale and rotation estimation is an important and challenging problem in visual object tracking. There have been proposed many sophisticated trackers to track the location of a target accurately, but most of them do not take much attention to the scale and rotation estimation. Inspired by the success of the correlation filters in visual tracking, we proposed a novel scale-and-rotation correlation filter (SRCF) in the Fourier domain to realize the scale and rotation estimation. We thus constructed a tracker with this scale-and-rotation correlation filter and a corrected kernel correlation filter. Our tracker was tested on a full benchmark dataset consisting of 50 video in comparison with fifteen state-of-art trackers. Both the success plots and the precision plots show that our tracker achieved superior performance in real time.

Index Terms— object tracking, correlation filter, scale and rotation estimation, log-polar transform

1. INTRODUCTION

Visual object tracking is a classical problem in computer vision and pattern recognition. The task of tracking can be described as capturing the target region in each frame of an image sequence; it can be applied to the motion analysis, surveillance, intelligent traffic and human-computer inter-action et al [1]. In the past decades, visual object tracking has achieved much progress in the efforts of researchers all over the world, but it remains a very challenging problem to develop a robust method for complex scenes, especially when the target undergoes serious scale and rotation variations.

Generally, object tracking algorithms can be categorized as the generative methods [2, 3, 4, 5, 6, 7] and the discriminative methods [8, 9, 10, 11, 12, 13, 14, 15, 16, 17] according to their appearance models and search strategies. The generative tracking methods usually represent the target object using an appearance model and search the object region by mining the reconstruction error. The discriminative algorithms pose the tracking problem as a binary classification task with local search and determine the decision boundary for discriminating the target object from the background. Recently, the discriminative trackers have shown excellent tracking performance. Among them, the correlation filter based trackers [10, 12, 13, 15] have achieved a huge success for their simplicity and effectiveness. They take advantage of the fact that the convolution of two patches in the space domain is equivalent to a point-wise product in the Fourier domain, so that they can conduct tracking using only a light computational power.

The MOSSE tracker [10] finds an adaptive correlation filter by minimizing the sum of squared errors between the expecting outputs and the actual outputs. It is robust to variations in lighting, pose and non-rigid deformations while operating at hundreds of frames per second. By employing the kernel trick and circulatory matrix on the correlation filter, the KCF tracker [12] provides best performance in tracking precision on the recent benchmark dataset OTB2013 [1] in real-time. Despite their great success, the above tracking methods can not deal with the scale and rotation changes of the target. To overcome this limitation, the DSST [15] was proposed for estimating the target scale by training a classifier on a scale pyramid and reached top performance in the VOT2014 challenge. However, the above methods rarely of take the rotation of the target into account, which is essential to track the target accurately.

In this paper, we will concentrate on this challenging problem of scale and rotation estimation in visual object tracking. Inspired by FFT-based image registration technique [18], it can match the images that are translated, scaled and rotated with respect to one another. The scale and rotation variations of an image in the spacial domain can be expressed as pure translation by mapping its Fourier magnitude spectra to the log-polar plane. Based on this idea, we realized the scale and rotation estimation by constructing a scale-and-rotation correlation filter in the log-polar domain. Since the Fourier magnitude spectra is invariant to translation, it is reasonable to estimate the scale and rotation before the translation of the target. The KCF and MOSSE perform correlation and updating using image patches of fixed scale and rotation angle, and they are likely to fail when the target changes its size or rotates in the image plane. Therefore, we proposed conducting the correlation and updating for the KCF tracker with resized target templates, named as the cor-

This work is supported by the National Natural Science Foundation of China Project No.61173110

rected kernel correlation filter. We construct our tracker by estimating the scale and rotation with our scale-and-rotation correlation filter and the translation with the corrected kernel correlation filter. Our tracker has shown a superior performance compared to the state-of-art trackers by carrying out experiment on the OTB2013 dateset.

The rest of the paper is organized as follows: in the next section, the scale-and-rotation correlation filter and our proposed tracker are described in detail. The experimental results are presented in Section 3, and finally some conclusions are drawn in Section 4.

2. OUR APPROACH

2.1. Scale-and-rotation correlation filter

In order to estimate the scale factor and rotation angle of the target, we proposed a correlation filter based scale and rotation estimation approach. We will firstly show the relationship between two translated, scaled or rotated target templates, and then present our scale-and-rotation correlation filter.

Let $f_2(x, y)$ be a replica of an image patch $f_1(x, y)$ with translation (x_0, y_0), scale factor a and rotation angle θ_0, that is

$$f_2(x,y) = f_1(\frac{xcos(\theta_0) + ysin(\theta_0) - x_0}{a}, \frac{-xsin(\theta_0) + ycos(\theta_0) - y_0}{a}) \quad (1)$$

Their Fourier transforms are related by

$$F_2(\xi,\eta) = \frac{e^{-j2\pi(\xi x_0 + \eta y_0)}}{a^2} \times F_1(\frac{\xi cos(\theta_0) + \eta sin(\theta_0)}{a}, \frac{-\xi sin(\theta_0) + \eta cos(\theta_0)}{a}) \quad (2)$$

Letting M_1 and M_2 be their Fourier magnitude spectra, and ignoring the multiplication factor $1/a^2$, we have

$$M_2(\xi,\eta) = M_1(\frac{\xi cos(\theta_0) + \eta sin(\theta_0)}{a}, \frac{-\xi sin(\theta_0) + \eta cos(\theta_0)}{a}) \quad (3)$$

Their Fourier magnitude spectra in the log-polar representation are related by (4).

$$M_2(log\rho, \theta) = M_1(log\rho - loga, \theta - \theta_0) \quad (4)$$

We can write (4) as

$$M_2(\phi, \theta) = M_1(\phi - d, \theta - \theta_0) \quad (5)$$

where $\phi = log\rho$ $d = loga$.

From (5), we know that the scaling and rotation between two images can be expressed as pure translations by converting their Fourier magnitude spectra from the Cartesian coordinates into the log-polar coordinates. Hence, we construct a correlation filter to estimate the translation in the log-polar domain. Once we get their translation in the log-polar domain, its scale factor and rotation angle will be found.

To start, it needs a set of training images $f_i(i = 1, 2, ..., n)$ and expecting training outputs $g_i(i = 1, 2, ..., n)$. For the training images f_i, they are rotated and scaled target templates. In our case, each g_i is generated from the ground truth such that it has a 2D Gaussian shaped peak centered at the target patch with the translation in the log-polar domain. First of all, we calculate the Fourier magnitude spectra $M_i(\xi, \eta)$ of all the training images f_i. Then, their log-polar representations $M_i(\phi, \theta)$ are obtained by the logarithmic polar conversion. Just like the MOSSE tracker, we construct a correlation filter HSR to estimate the translation of target patches in the log-polar coordinates. Let $\hat{M}_i$ (we use the $\hat{}$ denoting the Fourier transform in following content) and G_i denote the Fourier transform of $M_i(\phi, \theta)$ and g_i, the filter can be obtained by solving following equation (6).

$$min_{HSR^*} \sum_i |\hat{M}_i \odot HSR^* - G_i|^2 \quad (6)$$

The formulation (6) has an identical form of [10], but they have several differences. In our case, the training samples $M_i(\phi, \theta)$ are the log-polar representation of the training images' Fourier magnitude spectra, and training images f_i are scaled and rotated patches of target templates. On the other hand, the desired correlation output g_i is constructed with translations in the log-polar domain.

By solving for HSR^* with the method presented in [10], a closed form expression of the scale-and-rotation correlation filter is found:

$$HSR^* = \frac{\sum_i G_i \odot \hat{M}_i^*}{\sum_i \hat{M}_i \odot \hat{M}_i^* + \lambda} \quad (7)$$

where λ is the regularization parameter to avoid division-by-zero.

Given a detection patch z in the new frame cropped at previous position, the correlation score y can be computed by $y = F^{-1}(HSR_t^* \odot \hat{Z})$, where Z is the log-polar representation of its Fourier magnitude spectra. Once we get the translation (d_0, θ_0) that maximizes the correlation score, the scale factor can be obtained by $a = e^{d_0}$ and the rotation angel is θ_0. The training and evaluation steps are performed efficiently using the FFT, so that our scale-and-rotation correlation is very efficient. We construct our filter to conduct tracking according to above approach at the first frame, and update our filter in the following manner:

$$HSR_t^* = \frac{A_t}{B_t} \quad (8)$$

$$A_t = \eta G_t \odot \hat{M}_t^* + (1-\eta)A_{t-1} \tag{9}$$

$$B_t = \eta \hat{M}_t \odot \hat{M}_t^* + (1-\eta)B_{t-1} \tag{10}$$

where η is the learning rate, and A_1 and B_1 are initialized to the numerator and denominator of HSR^* respectively in the first frame.

2.2. Corrected kernel correlation filter

In [12], the author derived a kernel correlation filter HT from the kernel regression to estimate the location of a target, which has the form:

$$HT = \frac{G}{\hat{k^{xx}} + \lambda} \tag{11}$$

where k^{xx} denotes the kernel correlation of the training patch x and G denotes the Fourier transform of expecting output g. In current frame, the detection response y of detection patch z can be found out by (12). For the details and derivation, you can get from the [12].

$$y = F^{-1}(\hat{k^{xz}} \odot HT) \tag{12}$$

where k^{xz} denotes the kernel correlation between x and z. In the process of tracking, the kernel correlation filter is updated in the following manner:

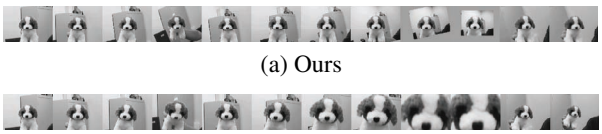
$$HT_t = \eta HT + (1-\eta)HT_{t-1} \tag{13}$$

where η is the learning rate, HT is the computed kernel correlation filer in current frame.

Among the frames a target may undergo scale change and rotation variation besides the translation. We firstly estimate the scale factor and rotation angle, and then resize the target to its original size using the estimated scale and rotation by the bilinear interpolation. The resized target in the detection patch of our approach is translated only with each other between adjacent frames, while the target in the detection patch of the KCF is a result of translation, scaling and rotation. It will be accurate to estimate the translation with the resized target patch, and the kernel correlation filter will not be wrapped with the improper patch when updating. We choose some representative detection patches in the *Dog1* sequence to illustration, that can be seen from **Fig.1**.

2.3. Proposed tracking method

Our algorithm firstly estimates the scale and rotation with our scale-and-rotation correlation filter, and subsequently gets the translation of the target in the current frame using the corrected kernel correlation filter, which is briefly outlined in **Algorithm SRCF**.



(a) Ours

(b) KCF

Fig. 1. Representative patches for correlation and updating of the *Dog1* sequence. Patches used in our model are shown in (a), the KCF in (b).

Algorithm SRCF: iteration at time t

Input:

Current image I_t.

Previous target position P_{t-1}, scale factor s_{t-1} and rotation angle θ_{t-1}.

Scale-and-rotation correlation filter HSR_{t-1}.

Corrected kernel correlation filter HT_{t-1}.

Setp1: Estimate the scale s_t and rotation θ_t at the current frame from P_{t-1}, s_{t-1} and θ_{t-1} using the scale-and-rotation correlation filter HSR_{t-1}.

Setp2: Estimate the target position P_t in the current frame from P_{t-1}, s_t and θ_t using the corrected kernel correlation filter HT_{t-1}.

Step3: Update the scale-and-rotation correlation filter according Eqs.(8), (9) and (10), and the corrected kernel correlation filter according to Eq.(13), using the current frame at P_t with s_t and θ_t.

Output:

Target position P_t, scale factor s_t and rotation angle θ_t.

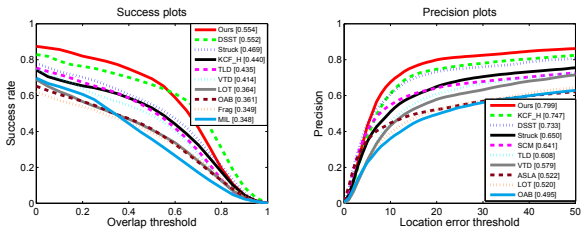
Updated scale-and-rotation correlation filter HSR_t and corrected kernel correlation filter HT_t.

3. EXPERIMENT

3.1. Experiment setup

We implemented our tracker according to **Algorithm SRCF** by MATLAB without optimization, and evaluated our tracker on the recent benchmark dataset OTB2013 [1] which contains 50 videos. All the experiments are conducted on an Inter i5-2400 CPU (3.1GHz) PC with 8 GB memory.

We choose the success plots and precision plots as our evaluation criteria [1]. The success plots measure the ratio of the successful frames whose overlap scores S are larger than a given threshold τ_0. Given the tracked bounding box R_t and the ground truth bounding box R_a, the overlap score is defined as $S = \frac{|R_t \cap R_a|}{|R_t \cup R_a|}$, where $\cap$ and $\cup$ represent the intersection and union of two regions respectively, and $|\cdot|$ denotes the number of pixels in a region. We use the AUC (area under curve) of a success plot to rank the algorithms. The precision is defined as the frame average Euclidean distance between



(a) Success plots (b) Precision plots

Fig. 2. Success and precision plots for all 50 sequences in the dataset OTB2013. The performance score for each of the trackers is shown in the legend. In each of the figures, the top 10 trackers are presented for clarity.

the center locations of the tracked targets and the manually labeled ground truth. The precision plot shows the percentage of frames whose estimated location is within the given threshold distance from the ground truth. As the representative precision score for each tracker we use the score for the $threshold = 20$ pixels [9].

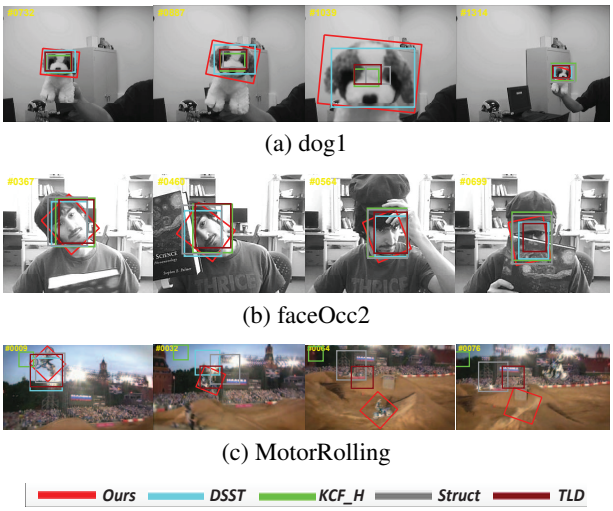
3.2. Quantitative experiment

We provide the quantitative and qualitative analysis of our approach SRCF in comparison with the existing state-of-art tracking method, including KCF [12], MOSSE [10], DSST [15], L1APG [2], CT [14], Struct [16], TLD [17], MIL [9], SCM [6], OAB [11], ASLA [7], Frag [8], LOT [5], IVT [3], and VTD [4]. Their overall performance on all the 50 videos in the dataset OTB2013 is summarized in the success plots and precision plots, as showed in Fig.2(a) and Fig.2(b).

From **Fig.2**, we see that our tracker gains the best performance both in the success plots and in the precision plots.Because the DSST configured a scale filter on the correlation filter based tracker, it gets the first place in the VOT 2014 challenge, and it gets also better performance on the success plots in our experiments. Our tracker configures a scale-and-rotation correlation filter, and gets a better performance than DSST. Moreover, our tracker can estimate the translation more accurately by the corrected kernel correlation filter, and our tracker does yield a noticeable increase than the KCF in the precision plots. At the same time, our tracker efficiently processes all the videos at an average fps of 43.

3.3. Qualitative experiment

In this part, we show the tracking results on several representative sequences that have serious scale and rotation changes. For clarity, only the results of the top 5 trackers in the success plots are showed in **Fig.3**. For the *dog1* sequence in **Fig.3**(a), the target suffers significant scale changes and slight rotation. Because our tracker configures a scale-and-rotation correlation filter, it estimates both the scale and rotation of the target



(a) dog1

(b) faceOcc2

(c) MotorRolling

Fig. 3. The tracking results of the top 5 trackers including ours in the success plots on three representative sequences.

accurately, while the other trackers fail. In **Fig.3**(b), the target has in-plane rotations, scale changes and occlusions, in which most of the trackers can determine the translation of the target, but none of them captures the rotation of target except our tracker. Considering the *MotorRolling* sequence in **Fig.3**(c), the target (biker and his motorbike) undergoes several deformations as the biker performs an acrobatics routine, which leads to significant changes in the scale as well as the rotation angle of the bounding box. Our tracker estimates the scale, rotation and translation accurately in the first 70 frames, while all the others loss the target at the very beginning. Our tracker also losses the target around the 70th frame for rapid motion and heavy scale change of the target.

4. CONCLUSIONS

In this paper, we proposed a scale-and-rotation correlation filter to estimate the scale factor and rotation angle of the target for robust visual object tracking. The challenging problem of scale and rotation estimation could be addressed by our proposed filter. The experimental results demonstrate that our strategy of correcting the kernel correlation filter improves the performance of the corresponding tracker. Both the success plots and the precision plots on the benchmark dataset OTB2013 demonstrate the state-of-art performance of our tracker with a high efficiency. Moreover, our scale-and-rotation correlation filter can be incorporated into any tracking framework easily. In the future, we will focus on the estimation of the six parameters in the affine transformation in order to adapt to the more general distortion of the target, as well as incorporating the kernel trick into the scale-and-rotation correlation filter.

5. REFERENCES

[1] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Computer vision and pattern recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2411–2418.

[2] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1830–1837.

[3] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[4] Junseok Kwon and Kyoung Mu Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1269–1276.

[5] Shaul Oron, Aharon Bar-Hillel, Dan Levi, and Shai Avidan, "Locally orderless tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1940–1947.

[6] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang, "Robust object tracking via sparsity-based collaborative model," in *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1838–1845.

[7] Xu Jia, Huchuan Lu, and Ming-Hsuan Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1822–1829.

[8] Amit Adam, Ehud Rivlin, and Ilan Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 798–805.

[9] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 983–990.

[10] David S Bolme, J Ross Beveridge, Bruce Draper, Yui Man Lui, et al., "Visual object tracking using adaptive correlation filters," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2544–2550.

[11] Helmut Grabner, Michael Grabner, and Horst Bischof, "Real-time tracking via on-line boosting.," in *BMVC*, 2006, vol. 1, p. 6.

[12] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 3, pp. 583–596, 2015.

[13] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming-Hsuan Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Computer Vision–ECCV 2014*, pp. 127–141. Springer, 2014.

[14] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang, "Real-time compressive tracking," in *Computer Vision–ECCV 2012*, pp. 864–877. Springer, 2012.

[15] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[16] Sam Hare, Amir Saffari, and Philip HS Torr, "Struck: Structured output tracking with kernels," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 263–270.

[17] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Tracking-learning-detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1409–1422, 2012.

[18] B Srinivasa Reddy and Biswanath N Chatterji, "An fft-based technique for translation, rotation, and scale-invariant image registration," *IEEE transactions on image processing*, vol. 5, no. 8, pp. 1266–1271, 1996.