Accurate Aspect Ratio Estimation in Scale Adaptive Tracking

Weikang Lu\textsuperscript{1,2,a}, Ang Su\textsuperscript{1,2,b}, Wenlong Zhang\textsuperscript{1,2} and Yang Shang\textsuperscript{1,2}

\textsuperscript{1}College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, Hunan, 410073, China
\textsuperscript{2}Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, Changsha, Hunan, 410073, China
\textsuperscript{a}Corresponding author’s e-mail: luweikang@nudt.edu.cn
\textsuperscript{b}E-mail: suang2008@126.com

Abstract. In visual object tracking, robust scale estimation is always a challenging problem. Because of the fast movement of the target, the background, the relative distance and relative azimuth angle between the target and the background usually change greatly. However, traditional correlation filtering tracking method cannot track the target accurately or even lose the target when the target scale and aspect ratio change. To address these problems, this paper adds two independent filters for the length and width scale estimation to complete the adaptive tracking on the basis of the translation filter. The experimental results in the public data set show that: compared with the original trackers, our method improves the tracking accuracy significantly, and meets the needs of real-time target tracking.

1. Introduction

Visual object tracking has always been a challenging problem in computer vision, which is also a research hotspot of scholars at home and abroad. With the wide application of video surveillance and human-computer interaction in people’s lives, and unmanned detection, deep space detection and precision strike are also widely used in the field of national defense and military\cite{1}, more and more moving targets, which have a series of deformation, occlusion, rapid movement, scale change and other issues in complex scenes pose a higher challenge to the precise tracking of moving targets.

Among the traditional target tracking approaches\cite{2,3}, correlation filters as the representative of the target tracking approach, with its high efficiency and good accuracy, has been widely concerned by researchers at home and abroad for its excellent performance. With the help of Fourier transform, the correlation filters transform the image matrix operation in the spatial domain into the point multiplication operation in the frequency domain, which greatly reduces the amount of calculation. Henriques\cite{4} et al. introduced kernel function and HOG features into correlation filter and proposed kernel correlation filter (KCF), which greatly improved the accuracy of the tracker, but still could not solve the problem of target scale change. In order to solve the problem of scale change, Martin\cite{5} proposed the DSST. In this paper, the position estimation and scale estimation are calculated separately. On the basis of the position estimation filter, an additional scale estimation filter is trained to update the target position and estimate the target scale, which effectively improves the scale adaptive ability of the tracker. However, when the target attitude changes, the aspect ratio of the target also changes. At this time, the initially given target frame obviously cannot adapt to this change. Sometimes a part of the target goes beyond the target frame, and only the local feature information of the target can be obtained.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
by tracker. On the contrary, sometimes there is a lot of background information in the target frame, and learning too much background information will cause the drift or failure of tracking.

In order to solve the above problems and improve the stability and comprehensive performance of the tracker, on the basis of correlation filtering, this paper adopts the target feature expression of Gray features and HOG features, which improves the target expression ability of the model. Aiming at the problem of scale change in the direction of length and width, two independent filters are designed to estimate the change of target length and width, which effectively improves the scale adaptive performance.

2. Principle of Correlation Filters

The basic principle of correlation filtering is to extract the features of the initial image patch \( f \) in the initial frame, after that, the filter \( h \) is trained so that the correlation of \( f \) and \( h \) response output \( g \) is a Gaussian distribution map, and the correlation calculation is accelerated by Fourier transform, the peak position of output map is the centre of the image.

When it comes to correlation filtering, it has to be said that the article [6], published in CVPR in 2010, brings correlation filtering into the field of target tracking for the first time. Through fast Fourier transform, the correlation operation is transferred to the frequency domain, and the correlation operation in the spatial domain is transformed into point multiplication in the frequency domain, which greatly improves the calculation speed. In this paper, the least output sum of squares error (MOSSE) is proposed, which is to find a filter to get the minimize output sum of squares error between the correlation result and the expected output. It improves the over-fitting in traditional trackers, improves the stability of tracking, and lays the foundation for a series of excellent tracker such as CSK, DSST, KCF and so on.

2.1. Basic principle of kernelized correlation filters

The paper [4] published in CVPR in 2015, is also a milestone in the development of correlation filtering. KCF introduces kernel function and multi-channel features based on Mosse, and generates multiple training samples \((x_i, y_i)\) by cyclic shifts method. Where \(x_i\) is the sample characteristic column vector, \(y_i\) is the label scalar. In order to prevent over-fitting, the regularization penalty term parameter \(\lambda\) is added, and the loss function is established by the ridge regression method:

\[
loss = \min_w \sum_i (f(x_i) - y_i)^2 + \lambda \|w\|
\]

(1)

Where \(f(x_i) = \omega^T x_i\), and when the derivative of \(\omega\) is 0, a closed solution can be obtained:

\[
\omega = (X^TX + \lambda I)^{-1}X^TY
\]

(2)

Where \(X^T = (x_1, x_2, \cdots, x_n)\), a cyclic matrix constructed by the feature of sample \(X\). In this paper, the kernel function of SVM is used to transform the nonlinear problem into a linear problem in high-dimensional space:

\[
\omega = \sum_i \alpha_i \varphi(x_i)
\]

(3)

Where, \(\varphi(x_i)\) is the mapping from low dimensional space to high dimensional space. Hence, the closed form solution of the classifier is obtained as follows:

\[
\alpha = (K + \lambda I)^{-1}Y
\]

(4)

Where \(K\) is the kernel correlation matrix of training samples: \(K_{ij} = \kappa(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle\). Where \(\varphi(x_i), \varphi(x_j)\) is the eigenvector of the training sample. Moreover, under the condition of partial kernel function, the kernel correlation matrix can still be guaranteed to be cyclic matrix. Hence after the Fourier transform, we can take advantage of the fact that the cyclic matrix can be diagonalized by the Fourier transform matrix:

\[
\hat{\alpha} = \frac{\gamma}{k^{xx+}\lambda}
\]

(5)

Where \(\hat{\alpha}\) is the result after Fourier transform. After the training, it is necessary to detect target in the new frame. After calculating the correlation between the filter and the candidate area of the new frame, the one with the largest response value is the location of the target:

\[
f(z) = K^2 \alpha
\]

(6)
Where, $K^z = \kappa(z_i, x_i) = \langle \varphi(z_i), \varphi(x_i) \rangle$, and $z_i$ is the feature vector of candidate samples.

2.2. Basic principle of discriminative scale space tracking

Discriminative scale space tracker (DSST) is an improvement based on Mosse which proposed a three-dimensional scale adaptive filter. The main idea of DSST is to estimate the position of the target first, and then estimate the scale of the target in a new frame. Hence the position filter and scale filter are independent of each other. Note that the principle of DSST is based on the assumption that the scale change of the target in two adjacent frames is often less than the position change of the target. Therefore, the scale information of the target can be determined more accurately by using the scale filter based on determining the position of the target. In addition, multi-channel HOG feature is used to make the description of target more reliable. And we can obtain the loss function:

$$
\text{loss} = \sum_{l=1}^{l} h^l \star f^l - g^l + \lambda \sum_{l=1}^{l} \| h^l \|^2
$$

(7)

Where $l$ is dimensions of HOG features and the star $\star$ denotes circular correlation. In order to achieve scale adaptation, DSST adds an additional scale filter to form a three-dimensional filter of size $(M \times N \times S)$, where $M \times N$ is the length and width of the target, $S$ is the number of scales. During training, the scale filter takes the target image as the centre, intercepts $S$ candidate frames with different scales, and obtains the feature vector $f$ at each scale. Then a $lS$ Gaussian distribution response graph $G$ is constructed, and we can use formula (7) to obtain filter $H$. During detection, the scale with the largest response value is the current scale of the target. At last, the obtained target scale is returned to the position filter to achieve more accurate position estimation.

3. Our Approach

3.1. Algorithm procedure

On the basis of KCF, the idea of DSST is used for reference, and scale filter is added to realize the scale adaptive ability of KCF. However, when the aspect ratio of the target changes sharply, it can achieve the requirement of accurate estimation only depending on the scale change. Therefore, this paper designs two independent filters to estimate the length and width of the target. The KCF is used to obtain the maximum response position of the target, and then the $X$ filter and the $Y$ filter are used to estimate the change of the target scale in the $X$ direction and the $Y$ direction, respectively, to determine the aspect ratio of the target, so as to realize the scale adaptation of the change of the aspect ratio of the target.

3.2. Location estimation

We use the kernel function proposed in KCF to speed up the calculation of the position filter and Gaussian kernel function is selected:

$$
K_{xx'} = \exp (-\frac{1}{\sigma^2}(\|x\|^2 + \|x'\|^2 - 2(\mathcal{F}^{-1}(\hat{x} \odot \hat{x}'))^T)
$$

(8)

Where $\sigma$ is the kernel function parameter, $\mathcal{O}^*$ represents the complex frequency domain conjugate, and $\mathcal{F}^{-1}$ is the inverse Fourier transform.

The kernel function maps the low-dimensional feature space to the high-dimensional space, making the features linearly separable. At the same time, in order to reduce the amount of computation, the image patch with more than 10000 pixels in the first frame is selected for $S$-times down sampling, so that the initial training image patch is less than 10000 pixels.

Moreover, in the process of image feature extraction, the Gray features of the target is added. In this paper, the HOG features and the Gray features of the image are fused to achieve a better description of the target.

3.3. Aspect ratio estimation

We add a scale estimation filter on the basis of KCF. After determining the target position, we keep the scale in the $y$ direction unchanged, and only change the scale in the $x$ direction. After obtaining the maximum response in the $x$ direction, tracker bring the scale estimation in the $x$ direction into $y$ filter
and get the scale in the y direction. Finally, the scales in the two directions are both obtained. The scale estimations in the x and y directions are derived by MOSSE. In this paper, we select the scale factor $a=1.02$, and the scale space series $S=33$, construct a $1*S$ Gaussian distribution correlation response graph $g$, and use the loss function to get determine filter $h$. Training the target area of each frame can get an optimal filter, but the calculation is computationally expensive, so only one frame of training sample is solved in engineering. The filter $h$ is determined by the loss function.

$$H^l = \frac{\sum_{i=1}^{d} F_k^l F_k^{k+1}}{F_k^{k+1}}$$

(9)

Then update the numerator and denominator $A_i^l$, $B_i^l$ of the relevant filter to:

$$A_i^l = (1 - \eta)A_{i-1}^l + \eta G_i F_i^l$$

(10)

$$B_i^l = (1 - \eta)B_{i-1}^l + \eta \sum_{k=1}^{d} F_k^{k+1}$$

(11)

Where $\eta$ is a learning rate parameter. Use formula (13) to calculate the relevant response value $y$ at the target area $z$, and the maximum relevant response value obtained is the new state of the target:

$$y = \mathcal{F}^{-1} \left\{ \frac{\sum_{i=1}^{d} A_i^l F_i^l}{B + \lambda} \right\}$$

(12)

The scale change is transferred to the position filter of the next frame to realize the scale adaptive tracking of aspect ratio change.

### 3.4. Training target extraction

In KCF, the target box is fixed at the initial moment, and the training sample area is $(1 + padding) * (M, N)$, centered on the target. If the training sample area is not updated, with the change of scale, the filter may learn more background information or focus too much on the local information of the target, and thus update the filter template by mistake, resulting in tracking drift or even failure. Therefore, we get the maximum response value of the latest frame through aspect ratio estimation, and return it to the position filter. According to the scale change of the last frame, we update the training sample area in real time. Therefore, our tracker can extract the target features accurately, and improve the robustness. The new training sample area is as follows:

$$X = (1 + padding) * (M * Scale_x, N * Scale_y)$$

(13)

Where $X$ is the training sample area, and $Scale_x$ is the scale change in the x direction, $Scale_y$ is the scale change in the y direction.

### 4. Experiments

In order to evaluate the effectiveness of the tracker, we compare the approach proposed in this paper with the KCF, DSST and other excellent trackers, and select the public data set OTB-50 [7] video sequence frames and part of the VOT [8] video sequence for testing.

All experiments were completed under MATLAB R2019a, 2.30 GHz CPU PC, and Window 10 operating system.

### 4.1. Quantitative comparison

For target tracking, precision and success rate are undoubtedly the most intuitive and effective evaluation criteria.

Precision is defined as the ratio of the number of tracking frames which less than the centre position error threshold to the total number of frames in the video sequence. The centre position error is defined as the mean Euclidean distance between the centre of the target tracking position and the true centre value. In this paper, the change range of centre position error threshold is set to 0-50 pixels.

The success rate is defined as the ratio of the number of tracking frames that reach the overlap rate threshold to the total number of frames in the video sequence. $S$ is defined as the overlap ratio between the tracking result and the real result in t frame:

$$S = \frac{\Phi_t \cap \Phi_a}{\Phi_t \cup \Phi_a}$$

(14)
Where $\varphi_t$ is the target tracking results, and $\varphi_g$ is the ground truth in frame $t$. Changing the overlap rate in a certain range, we can get the success rate of target tracking under different thresholds. In this paper, $S=0.5$.

Compared with 4 state-of-the-art trackers: DSST、KCF、Struck[9] and CSK[10], the comparison of the success rate and accuracy on the otb-50 data set is shown in figure 1:

![Figure 1. Precision and success plots over all the OTB-50 sequences.](image1)

The comparison results of the trackers can be obtained by selecting the test sequence with scale estimation, as shown in Figure 2:

![Figure 2. Precision and success plots over sequences with significant scale variation.](image2)

As can be seen from Figure 1, the approach proposed in this paper has a certain improvement in tracking success rate, accuracy and robustness compared with other trackers. In Figure 2, we can find that our method is better than other trackers in the data set with scale estimation. Because our method can accurately estimate the aspect ratio scale change of the target, and fuse the Gray features and HOG features of the target. Therefore, it has good robustness in dealing with the target rotation, illumination change and fast movement.

Table 1 shows a comparison of the average frame rate to other trackers on the data set.

| Trackers | KCF  | DSST | Stuck | CSK  | Ours  |
|----------|------|------|-------|------|-------|
| FPS      | 222.0278 | 36.939 | 10.0582 | 405.1675 | 51.8603 |

It can be seen that, compared with KCF, the speed of our method is obviously reduced due to the increase of scale estimation filter, but it is better than a series of tracker such as DSST and Struck. The approach proposed in this paper not only ensures the tracking accuracy, but also meets the real-time performance of the tracking.
4.2. Qualitative comparison

Figures 3-5 show the comparison of tracking results of different video sequences. In order to distinguish trackers, we use different colors to represent the target tracking box (green is our method, red is KCF, blue is DSST).

Figure 3 shows the partial tracking results of the video sequence *bolt_1*. The video sequence contains the scale change of the aspect ratio of the tracking target, and there are similar target interference and rapid movement and deformation of the target. When the angle of camera changes, the athlete’s body turns from the front to the side and then to the back. In frame 23, due to the rapid movement and rapid deformation of the athlete’s body, the KCF drifts. In frame 220, when athlete’s body changes, both KCF and DSST can locate the target, but they cannot accurately estimate the change of aspect ratio scale of the target. Our method is better adapted to the changes of athlete’s body.

Figure 4 shows the partial tracking results of the video sequence *girl*. The video sequence contains the deformation of the target and the temporary occlusion of similar targets. In frame 120, when the girl's head rotates, both KCF and DSST appear a certain degree of drift. In frame 470, when the man and the woman briefly cross over, both KCF and DSST mistakenly identify the man as the tracking target. Our method with accurate ground scale estimation can get effective target training samples so that the trainer can successfully distinguish positive and negative samples to achieve accurate target tracking.

Figure 5 shows the partial tracking results of the video sequence *carscale_1*. In the 100th frame, due to KCF cannot estimate the scale and focuses on the local feature description of the target, the tracking has obvious error. Although DSST can better locate the target position and estimate the scale of the target, the aspect ratio of the target changes with the progress of the vehicle. And DSST drifts on frame 192. By contrast, the approach proposed in this paper accurately captures the target.

In order to verify the effectiveness of our method for aspect ratio scale estimation, we also select some VOT data set with obvious aspect ratio scale change for experiments. Here we provide two sequences to compare, namely *dinosaur* and *fish3*.
Figure 6 shows the partial tracking results of the video sequence *dinosaur* in VOT dataset, which contains the aspect ratio change, rotation and complex background of the object. In the 155th frame, due to the rotation of the dinosaur model, KCF and DSST cannot estimate the aspect ratio scale change of the model. There is a lot of background information in the tracking frame, which is used to update the classifier in the next frame. With the accumulation of classifier errors, the trackers get more and more background features, which eventually leads to the failure of tracking. In the 240th frame, KCF and DSST have lost the target. Because the approach proposed in this paper can accurately estimate the change of aspect ratio of the target, it can get more accurate target training samples. Therefore, compared with other tracker, our method has better robustness.

In Figure 7, a partial tracking result of the video sequence *fish3* is shown. Due to KCF cannot estimate the scale and the single feature, the result has a big deviation. Similarly, there is a huge error between the tracking box and the real size of the target in DSST. Our tracker catches the target effectively by estimating the scale of the aspect ratio of the target and using the fused features.

### Figure 6. *dinosaur* sequence frame 1, 155, 240

### Figure 7. *fish3* sequence frame 28, 105, 307

However, no one tracker can adapt to all scenarios. The proposed tracker struggles in the basketball sequences, where the basketball player is doing irregular sports, and then nearly does not increase in size. This is most likely due to our method is too sensitive to changes in scale so that estimates the scale change of the target erroneously.

## 5. Conclusions

In this paper, a robust aspect ratio scale adaptive tracking approach is proposed. Based on the KCF and the DSST, two independent filters are used to estimate the change of the target in the length and width direction, and then the scale change of each update is used to estimate the position of the next frame to obtain more realistic and effective training samples. The Gray features of the target, fused with HOG features can improve the target positioning accuracy and scale adaptive performance.

Experiments are performed on public challenging benchmark sequences, both quantitative and qualitative evaluations are performed to validate our approach. The proposed tracker has achieved good tracking results in target scale estimation and also met the real-time requirements. However, in complex scenes, the occlusion of similar objects will cause failure of target tracking. Therefore, in the face of occlusion and complex scenes to achieve the robustness of target tracking is the focus of the next step.
References

[1] Liu Y T, Wang K F, Wang F Y. (2017) Tracklet association-based visual object tracking: the state of the art and beyond. ACTA Automatica Sinica, 43(11): 1869-1885.

[2] Yin H P, Chen B, Chai Y, et al. (2016) Vision-based Object Detection and Tracking: A Review. ACTA AUTOMATICA SINICA, 42(10): 1466-1489.

[3] Henriques J F, Caseiro R, Martins P, et al. (2012) Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the 12th European Conference on Computer Vision. Florence. pp. 702-715.

[4] Henriques J F, Caseiro R, Martins P, et al. (2015) High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3): 583-596.

[5] Danelljan M, Häger G, Khan F, et al. (2014) Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference 2014. Nottingham. pp. 65.1-65.11.

[6] Bolme D S, Beveridge J R, Draper B A, et al. (2010) Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. California. pp. 2544-2550.

[7] Wu Y, Lim J, Yang M H. (2013) Online object tracking: a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition. Oregon. pp. 2411-2418.

[8] Kristan M, Leonardis A, Matas J, et al. (2017) The Visual Object Tracking VOT2017 Challenge Results. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW). Venice. pp. 1949-1972.

[9] Hare S, Saffari A, Torr P. (2011) Struck: Structured output tracking with kernels. In: IEEE International Conference on Computer Vision. Barcelona. pp. 263-270.

[10] Henriques J F, Rui C, Martins P, et al. (2012) Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In: Proceedings of the 12th European conference on Computer Vision. Florence. pp. 702-715.