Adaptive Scale Correlation Tracking based on SVM

Kang Yuan[1,a)], Da-peng Wei[2,b)]

[1]College of Computer Science and Technology, Chongqing University of Posts & Telecommunications, Chongqing 400065, China

[2]Chongqing institute of green and intelligent technology, Chinese Academy Of Sciences, Chongqing 400714, China

[a)]:1019801567@qq.com

Key words: Object tracking; SVM; Multi-scale; Correlation filter

Abstract. Although the correlation filter-based trackers achieve the competitive results both on accuracy and robustness, there is still a need to improve the overall tracking capability. Focusing on the issue that the correlation filter-based trackers algorithm has poor performance in handling scale-variant target and Occlusion, this paper presents a multi-scale correlation filter algorithm combined with SVM detector to solve the above problems. Firstly, by introducing the scale factor into the kernel matrix to improve the performance of correlation filter processing scale transform. Then we trained an online SVM detector to retrieve the target when the target is occluded, and adaptively adjust the learning rate of the model. By comparing with the other six outstanding tracking algorithm, experimental results show that the proposed approach could estimate the object state accurately and handle the object occlusion problem effectively.

INTRODUCTION

Visual tracking plays an important role in compute vision, with a wide range of applications in intelligent monitoring, intelligent transportation, human interaction, machine vision, robotics and many other fields[1]. Although great progress has been made in the past decade, the model-free tracking is still a tough problem due to illumination changes, partial occlusions, fast motions and background clutters.

Tracking-by-detection trackers[2,3,4] are very popular due to its high performance and efficiency. In recent years, the correlation filter has been introduced into the visual tracking application, and achieved a good tracking effect. Bolme et al.[5] propose to learn a minimum output sum of squared error(MOSSE) filter for visual tracking on gray-scale images, where the learned filter encodes target appearance with update on every frame. Henriques et al. [6] propose a circulant structure of tracking-by-detection with kernels (CSK) approach, which exploits the circular structure of adjacent subwindows in an image for quickly learning a kernelized regularized least squares classifier of the target appearance with dense sampling. Later Henriques et al.[7] propose a kernelized correlation filter (KCF) tracking algorithm, which is further improved by using HOG features. Danelljan et al. [8] propose an adaptive color attributes tracking approach, which exploits the color attributes of a target and learns an adaptive correlation filter by mapping multichannel features into a Gaussian kernel space. However, these approachs are limited to predict the position of the target and can't handle occlusion and scale prediction, affect the tracking accuracy to a great extent. Based on the traditional CSK algorithm, we update the scale of the tracker with a kernelized scale filter, which represent the object with kernel feature space and extend kernelized correlation filter with a scale factor. Then, the re-detection mechanism based on the SVM is introduced to solve the occlusion problem.

KERNELIZED CORRELATION FILTERS BASED TRACKING

The CSK tracker learns a regularized least squares(RLS) classifier of the target appearance from a single image patch, gets the kernelized correlation filter with using the circulant matrices and

kernel trick, and localizes the target in a new frame by finding the maximum response of the correlation filter. In this section, we briefly describe the CSK tracker.

A classifier is trained using a single grayscale image patch x of size $M \times N$ that is centred around the target. The tracker considers all cyclic shifts $x_{m,n}$, $(m,n) \in \{0,\ldots,M-1\} \times \{0,\ldots,N-1\}$ as the training examples for the classifier. These are labelled with a Gaussian function y, so that $y(m,n)$ is the label for $x_{m,n}$ .The classifier is trained by minimizing the cost function(1) over w.

$$f(x) = \min_{w} \sum_{m,n} \left| \left\langle \varphi(x_{m,n}), w \right\rangle - y(m,n) \right|^2 + \lambda \langle w, w \rangle \tag{1}$$

Here φ is the mapping to the Hilbert space induced by the kernel κ, The constant λ is a regularization parameter. According to the properties of cyclic matrix, The ridge regression has the close-form solution: $\hat{w}^* = \dfrac{\hat{x}^* \odot \hat{y}}{\hat{x}^* \odot \hat{x} + \lambda}$

where $\hat{x}$ denotes the DFT of x, and $\hat{x}^*$ denotes the complex-conjugate of $\hat{x}$. In the case of no-linear regression, the sample is mapped to the high-dimensional feature space, defining the inner product as $\left\langle \varphi(x), \varphi(x') \right\rangle = \kappa(x, x')$, kernel trick, $f(z) = w^T z = \sum_{i=1}^{n} \alpha_i \kappa(z, x_i)$ is applied to allow more powerful classifier. By making use of the circulant structure among $x_{m,n}$ and the convolution theorem, the solution to the ridge regression problem is:

$$\mathcal{F}(\alpha) = \hat{\alpha} = \frac{\mathcal{F}(y)}{\mathcal{F}(\kappa^{xx}) + \lambda} \tag{2}$$

Here α is dual space coefficients, κ^{xx} is defined as kernel correlation to calculate the similarity between image samples. The detection step is performed by first cropping out a gray scale patch z of size $M \times N$ in the new frame.The detection scores are calculated as

$$\hat{y} = \mathcal{F}^{-1}(\hat{\alpha} H) \tag{3}$$

Where H=$\mathcal{F}$(h) ,where h is the vector with elements $h(m,n) = \kappa(z_{m,n}, X)$,where X represents the target model learned from the previous frame and $z_{m,n}$ is the sample of the image patch z. The position where $\hat{y}$ get the maximum response is the output of the target in the new frame. The target center position of each frame is determined by constantly iterating the Eq.2-3. For more details, we refer to [6].

THE PROPOSED VISUAL TRACKING ALGORITHM

The traditional CSK tracker uses image patch sample to train the classifier model with a fixed size. It is difficult to deal with the scale changes in the process of the target movement, which eventually leads to the cumulative error in the classifier mode and target drift. In this paper, the extend kernel scale filter approach is used to improve the multi-scale tracking of the CSK algorithm, and propose an effective occlusion processing mechanism to deal with the tracking failure.

Scale Evaluation

Similar to DSST[9], In this paper, we study an independent one-dimensional kernel correlation filter to detect target scale changes. Different from DSST, which only used the original feature space as the object representation, we represent the object with kernel feature space and extend kernelized correlation filter with a scale factor. The kernelized correlation filter can integrate multi-channel features [9]，which can improve the ability of classifier to distinguish. In the process of tracking, a series of image patchs of different scales centerd around the target are constructed as the sample, in order to reduce the computational complexity and preserve the coherence of object representation in different scales, we resize the scale of current training sample to the initial scale of the first frame, so that the feature dimensions of the target filter are consistent throughout the tracking process. The multi-scale kernelized correlation tracking filter H can be represented as:

$$H = \frac{Y\Phi\left(\varphi(x)\right)}{K\left(\varphi(x),\varphi(x)\right)+\lambda} \tag{4}$$

Where x is the image sample patch, $\varphi(\cdot)$ represent image feature, $\Phi(\cdot)$ is a feature mapping function in the Fourier domain, $K(\cdot,\cdot)$ is the kernel matrix to compute the kernel correlation. After H is obtained, the scale of the target area z in the next frame can be estimated,extract the corresponding HOG feature for target area z, denoted as g, then take the scale s_i of the maximum value of the formula $\mathcal{F}^{-1}(\mathcal{F}(\mathrm{g})\odot H^*)$ to track the target.

Online Detection

We use the re-detection module to handle the occlusion and adaptively adjust the model learning rate. Training a linear SVM[10] classifier as a detector. In the first frame, the SVM is trained by using user's input bounding boxes that do not overlap with the target as negative examples. In the subsequent frame, computed features are convolved with SVM weights, and we detect the top-n confident bounding boxes $C_t = \{c_1,\ldots,c_n\}$, in this paper, $n=8$. Generally the target state $\overline{s}_t$ (i.e. the object position and scale) can be found by maximizing the correlation score. If the overlap rate between the state $\overline{s}_t$ and one of the detected candidate bounding boxes Can_t is larger than $\mathcal{T}$,we consider the state $\overline{s}_t$ as the correct target state s_t in the t-th frame; otherwise, the state $\overline{s}_t$ may be not correct, and then we take use of C_t.To be specific, for each detection candidate bounding box we use the kernelized correlation filter to obtain the maximum correlation score $\hat{y}_i$ and the correlation score of the preliminary target state $\overline{s}_t$ as all candidate score $\hat{y} = \{\hat{y}_1,\ldots\hat{y}_n,\hat{y}_{n+1}\}$. To preserve the spatial-temporal consistency structure in consecutive frames, we adjust all candidate scores with spatial Gaussian distribution, which is based on the spatial distance between the candidate bounding box center and the last estimated object center. Then the corresponding candidate state of the maximum candidate correlation score is found as the final object target state s_t .When the object is occluded, the inappropriate update of object appearance may lead to model drift. To deal with the problem, we adaptively adjust the learning rate. If the object is occluded, we reduce the learning rate; if else, keep the learning rate. We adjust the learning rate β as follows:

$$\beta = \begin{cases} 0.1*\beta_{init}, & if \quad \mathrm{T}_o < \mathcal{T} \\ \beta_{init}, & otherwise \end{cases} \tag{5}$$

Where β_{init} is the initialization value of the learning rate β and $\mathcal{T} = 0.05$.

EXPERIMENT

Experimental Settings

Our approach is implemented using Matlab and C language mixed preparation. In the experiment, the parameters in the algorithm are the same for all test videos. The standard deviation of the Gaussian function σ is 0.5, the cell size of the HOG feature is 4 × 4 pixels, the orientation bin number is 9, the translation learning factor β is 0.075, the regularization parameter λ is 0.01, the scale pool $S = \{0.985, 0.99, 0.995, 1, 1.005, 1.01, 1.015\}$, the scale learning factor β_s is 0.025. In this paper, we have compared ours algorithm with six excellent trackers on ten video sequences from the Visual Benchmark[11].

Performance Evaluation

In all experiments, we use center location error(CLE)、distance precision(DP)、success rate(SR) as a comprehensive evaluation index. CLE represents the Euclidean distance between the center of the tracking result and the center of the groundtruth annotation，DP indicates the ratio between video frames in which the CLE is less than a fixed threshold (take $20\,pixel$ in experiment) and test video frames. The success rate is defined as

$$score = \frac{area\left(R_t \cap R_{gt}\right)}{area(R_t \cup R_{gt})} \tag{6}$$

Where R_t is the tracking bounding box and R_{gt} is the ground truth bounding box, $area$ is the area of bounding box, S_n is the number of tracking successful. If the score is higher than 0.5 in one frame, the tracking results are considered as a success, and add one for S_n , where $SR = S_n / N$. The seven algorithms were tested on ten video sequences to obtain the mean CLE、$SR\ and\ DP$, as shown in Table1, Table 2 and Table 3.

Table 1. Mean center location error(pixel)

Sequence	David	Trellis	Soccer	Skating1	Singer1	Girl	CarScale	Dog1	Woman	Jogging-1	Mean
Ours	7.43	**4.62**	14.3	**13.6**	14.1	3.77	30	**4.81**	11.2	**3.53**	**13**
CSK	27.1	26.5	63.9	22.9	79.9	17.7	55.7	17.1	208	133	58.65
TLD	8.14	31.1	**4.34**	72.5	17.9	9.9	**28.8**	15.7	18.9	5.72	21.48
KCF	9.7	6.99	20	22	62.1	13	51.8	17	10.3	95.3	30.49
CT	22.2	52.7	98.4	147	74.4	18.2	63.7	22.7	115	93.2	68.58
DSST	**4.78**	6.09	21.6	19.5	**5.41**	7.92	34.7	7.71	8.93	107	20.45
Struck	35.4	18.4	73.1	84.1	74.1	**3.54**	74.2	20	**6.13**	61.5	41.5

Table2. distance precision (%)

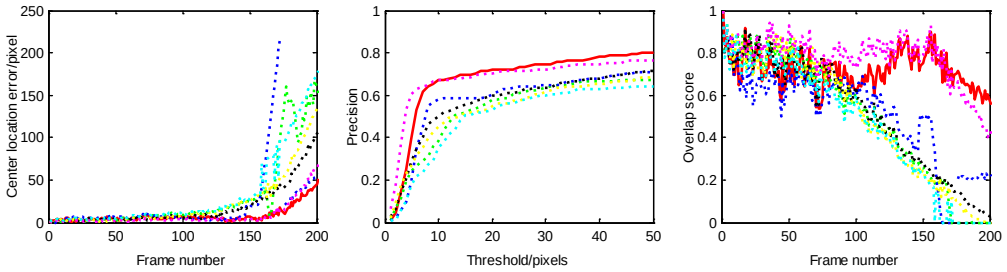
Sequence	David	Trellis	Soccer	Skating1	Singer1	Girl	CarScale	Dog1	Woman	Jogging-1	Mean
Ours	**100**	**100**	89.8	**68.5**	74.6	**100**	**71.8**	**100**	**93.8**	**97.4**	**82.23**
CSK	33.3	39.9	13.5	53.3	12	58.4	57.9	72.5	25	23.1	39.72
TLD	99.2	52.1	**100**	36	67	91.4	66.8	74.8	79.9	**97.4**	72.03
KCF	99.4	99.6	49	44.8	14.8	86.2	59.9	74	93.6	23.1	58.37
CT	41.8	15.8	21.9	5	15.4	65.8	54.4	69.6	20.4	23.1	31.25
DSST	**100**	**100**	73.5	59	**100**	93.6	70.2	98.9	93.6	23.1	80.71
Struck	36.7	78.7	16.3	37.8	16.8	100	53.2	70.8	96.3	24.4	48.92

Table 3. tracking success rates (%)

Sequence	David	Trellis	Soccer	Skating1	Singer1	Girl	CarScale	Dog1	Woman	Jogging-1	Mean
Ours	95.8	**98.4**	**44.9**	**85.5**	64.4	**98.2**	**80.6**	**99.9**	93.3	**96.7**	**82.53**
CSK	23.6	16.7	13.5	40.3	27.6	43.2	44.8	60.5	24.3	22.5	36.65
TLD	90.7	48	9.44	21	99.1	75	50.8	66.9	17.6	96.4	59.18
KCF	52.7	84	40.3	38.5	26.5	63.6	44.8	61.2	93	22.5	54.16
CT	23.8	17.8	19.9	5.75	20.8	14	41.7	46.4	15.4	22.5	28.15
DSST	**100**	97	28.3	49.3	**100**	37	76.2	99.8	93.3	22.5	75.18
Struck	23.6	77.5	15.6	29.3	29.9	98	41.3	61.6	**93.5**	21.8	51.87

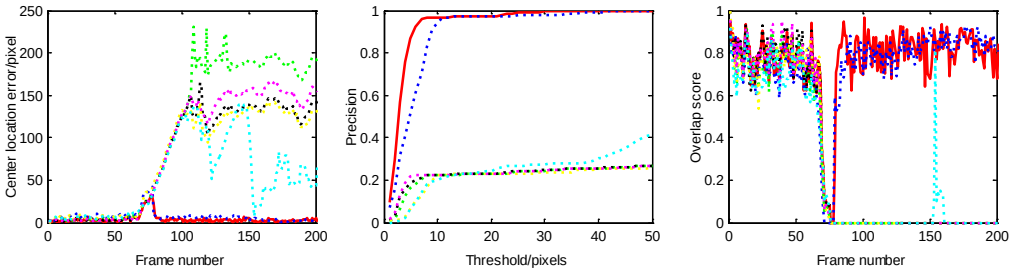
It can be seen from Table 1, Table 2 and Table 3 that the mean value of the mean center position error is reduced from the original 58.65 pixel to 13 pixel compared with the original CSK algorithm, and the mean value of the distance is increased from 39.72% to 82.23% , The success rate is increased from 36.65% to 82.53%. The quantitative analysis of these three evaluation criteria can prove that the tracking performance of our approach is better than CSK tracker. Compared with the other six tracker, the three evaluation values are also the best values, which proves that the tracking performance of our approach is obviously improved.

To describe the tracking results in detail, we give the center location error plots, the overlap plots, and the distance precision plots from partial experimental results which are shown in Figures 1 over 3 sequences for these trackers. From the figures 1,we can see that our tracker maintains a lower centre location error, a higher overlap score, and a higher distance precision in general. The above analysis implies that our approach performs more accurate and stable results than the other six trackers.



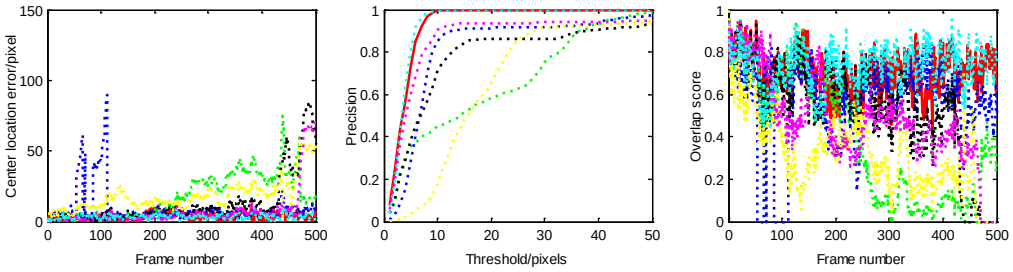
(a) CLE (b) DP (c) SR

(1)The Carscale sequence



(a) CLE (b) DP (c) SR

(2)The Jogging-1 sequence



(a) CLE (b) DP (c) SR

(3)The Girl sequence

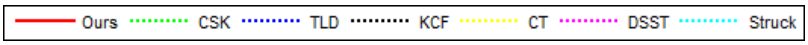
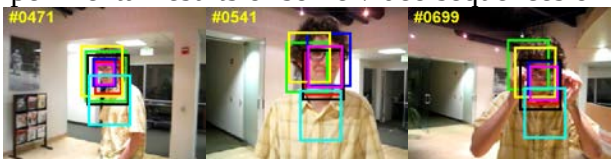


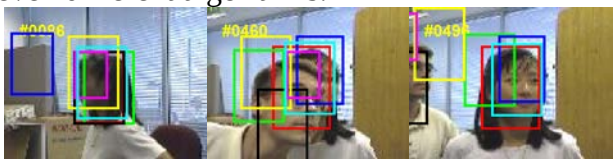
Fig.1. Quantitative comparison of the seven trackers with the center location error, distance precision and tracking success rates on the Carscale, the Jogging-1,and the Girl video sequences

Experimental Results

In order to verify the performance of the algorithm, this paper gives a comparison of the experimental results of some video sequences on seven different algorithms.



(1)David(#471,#541,#699)



(2)Girl(#86,#460,#496)

(3)Jogging-1(#21,#83,#177)

(4)Dog1(#192,#856,#1054)

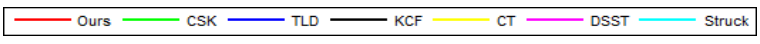


Fig.2. Part of the experimental comparison chart

In Figure 2(1), the person walks towards the moving camera, resulting in significant appearance variations due to the illumination and scale change. CT, KCF and our approach can successfully track the target in most frames of the David sequence. In Figure 2(2), the girl undergoes scale variation and partial occlusion which make the tracking more difficult, only our tracker is able to track the target successfully in most frames of this sequence. Figure 2(3) shows the Jogging-1 sequence with occlusion, deformation and rotation, when occlusion occurs, our approach and the TLD algorithm can still track the target accurately because of the re.-detection. Figure 2(4) shows the Dog1 sequence with scale and pose variation. All algorithms have an excellent result when there is no obvious variation in scale in the 192th frame; but our approach has a definite advantage when the scale is significant changed. Based on the above results, our approach has a good effect under the condition of scale change and occlusion.

CONCLUSION

In this paper, we proposes a tracking algorithm of multi-scale correlation filter based on the SVM. Our approach learns discriminative correlation filters for estimating the translation and scale variations of target objects effectively. we update the scale of the tracker with a kernelized scale filter, which represent the object with kernel feature space and extend kernelized correlation filter with a scale factor. We further develop a robust online detector using SVM to re-detect targets in case of tracking failure. The experimental results show that the tracking performance of our approach is higher than that of the original CSK algorithm, and it is obviously higher than the other six classical algorithms. It is suitable for moving target tracking in complex scenes with scale change and occlusion.

REFERENCES

[1] Fang J, Wang Q, Yuan Y. Part-Based Online Tracking With Geometry Constraint and Attention Selection[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2014, 24(5): 854-864.
[2] Zhang K, Zhang L, Yang M H. Real-Time Compressive Tracking[C]// European Conference on Computer Vision. Springer-Verlag, 2012:864-877.
[3] Kalal Z, Mikolajczyk K, Matas J. Tracking-Learning-Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(7): 1409-1422.
[4] Hare S, Golodetz S, Saffari A, et al. Struck: Structured Output Tracking with Kernels[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 23(5):263-270.
[5] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]// IEEE Conference on Computer Vision &Pattern Recognition. IEEE, 2010:2544-2550.
[6] Henriques J F, Rui C, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[M]// Computer Vision – ECCV 2012. Springer Berlin Heidelberg, 2012:702-715.
[7] Henriques J F, Rui C, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3):583-596.
[8] Danelljan M, Khan F S, Felsberg M, et al. Adaptive Color Attributes for Real-Time Visual Tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014:1090-1097.
[9] Danelljan M, Häger G, Khan F S, et al. Accurate Scale Estimation for Robust Visual Tracking[C]// British Machine Vision Conference. 2014:65.1-65.11.
[10] Chapelle O. Training a support vector machine in the primal[J]. Neural Computation, 2007, 19(5):1155-1178.
[11] Wu Y, Lim J, Yang M H. Online Object Tracking: A Benchmark[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2013: 2411-2418.