International Journal of Approximate Reasoning 154 (2023) 38–47

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Spatial-temporal single object tracking with three-way decision theory

Ziye Wang, Duoqian Miao *

Department of Computer Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 30 August 2022
Received in revised form 3 December 2022
Accepted 6 December 2022
Available online 9 December 2022

Keywords:
Object tracking
Three-way decision
Spatial-temporal network
Computer vision

ABSTRACT

Trackers based on Siamese network have achieved positive performance in recent days. However, most of the existing siamese single object trackers only consider the spatial information in the template which was given in the first frame of the video but do not extract the affluent temporal information. In this paper, we propose a novel tracking framework based on a spatial-temporal network. Specifically, we introduce three-way decision theory into object tracking to avoid interference from complex situations such as occlusions, fast motions, and non-rigid deformation. Furthermore, our proposed method can generate more precise tracking results due to the discriminative correlation filters (DCF). Extensive tests and comparisons with numerous competitive trackers on demanding large-scale benchmarks, including OTB-2015, GOT-10k, LaSOT and VOT2018, TrackingNet, demonstrate that our tracker outperforms many state-of-the-art real-time techniques while operating at 22 frames per second.

1. Introduction

Visual object tracking, which automatically tracks a predetermined target in a changing video sequence based on its initial annotation in the first frame, is a crucial computer vision task having applications in several research areas, including autonomous driving [1], video surveillance [3], and autonomous interactions [11]. Due to difficult elements such as occlusions, fast motions, and non-rigid deformations [21–24], how to recognize and locate the item precisely in changing circumstances is still a fundamental great gap in object tracking.

In recent years, discriminative correlation filters (DCF) based approaches for visual tracking projects, such as KCF [2], SAMF [4], LCT [41], SRDCF [7], and SCSTCF [62], have received a lot of attention. The majority of these methods use handcrafted features, which reduces their robustness and accuracy. Inspired by the advancements in deep learning techniques, significant attention has been paid to deep learning-based methods for visual tracking. Among these trackers, the most well-liked methodology approaches object tracking as a similarity matching problem between the target template and the search frames in an offline trained embedding space, called Siamese trackers [12–15,28,29,34,25,40,63]. However, most of these tracking methods typically don't update the template, or only take appearance features from the current frame, making it challenging to adjust to changes in appearance brought on by conclusions, fast motions, non-rigid deformations, etc.

To solve this problem, some trackers [5,6] designed complex template update algorithms as a solution to this complex issue. Therefore, these methods exhibit stronger robustness than Siamese trackers. However, online template updating con-

* Corresponding author.
E-mail address: dqmiaoi@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ijar.2022.12.003
0888-613X/

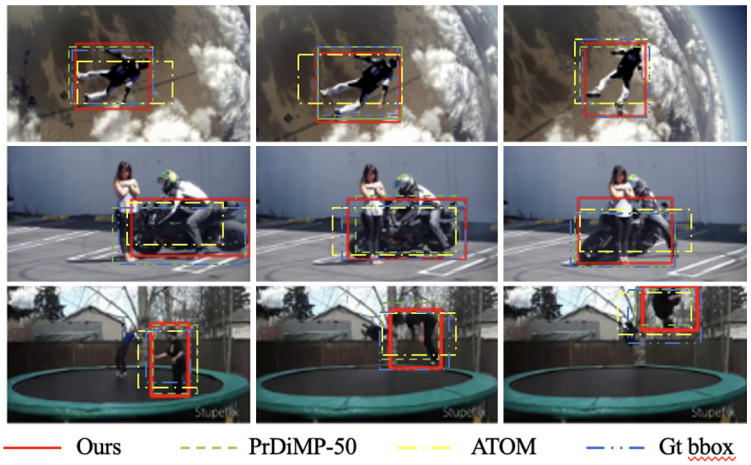

Fig. 1. The comparison of our proposed model with PrDiMP-50 and ATOM in challenging situations. Images show the predictions obtained using PrDiMP-50 and ATOM and our method, and the ground truth bounding box. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

sumes significantly more computational resources, which prevents trackers from real-time tracking. In addition, the majority of trackers now in use only consider the spatial information such as appearance features of the current frame and cannot benefit from temporal information like inter-frame information, which is a waste for video datasets containing rich historical temporal information. Therefore, considering the features of moving target change heavily in the complex situations, to maintain temporal continuity in a constantly changing environment, a novel method [76] was proposed to notice both spatial and temporal appearance information of the target. In multi-object tracking researching region, spatial-temporal mechanism also acquires more noticed like STAM [77] to help the tracker to be more robust to drift.

Features from diverse frames could provide us with different information for the same tracking object instance [56]. Although some trackers [56,78] utilize different temporal information to upgrade tracking performance, due to the lack of filtering of historical frames, the temporal information extracted from redundant frames and some frames in complex situation such as partial occlusion and deformation will interfere with the target tracking performance. Feature selection seeks to remove irrelevant or redundant characteristics while keeping relevant or informative features [8,80]. Note that the three-way decision problem of the information system with the fuzzy decision has received more attention [9,61] and samples could be well divided into positive region (POS), negative region (NEG), and border region (BND) by three-way decision theory. This motivates us to create a three-way decision based historical frame selection model to utilize both temporal information and spatial information, thus avoiding the influence of complex situations. In this paper, we develop a template-free single object tracking method. To this end, we formulate our spatial-temporal tracker which is combined with a historical frame selection model based on the three-way decision, thus avoiding the influence of changing circumstances such as occlusions, fast motions, and non-rigid deformations. Fig. 1 contains the ground truth bounding box (blue) and the tracking result obtained by PrDiMp-50 (green), ATOM (yellow) and our method (red). To evaluate the accuracy of the prediction of different methods, it can be considered that the closer to the ground truth bounding box, the better the tracking effect will be. Therefore, the Fig. 1 illustrates the advantage of our proposed method.

The contributions of this paper can be summarized in three folds.

We propose a novel end-to-end spatial-temporal single object tracking framework, which not only improves the tracking accuracy, but also improves the feature representation.

We propose a novel historical frame selection mechanism based on three-way decision theory, which enables our tracking method to select high-performance features in order to have greater robustness while tracking.

The proposed method is evaluated on 4 benchmarks: OTB-2015, TrackingNet, GOT-10k, LaSOT and VOT2018, while running in almost real-time at 22 FPS, which demonstrates the superiority of our framework.

2. Related works

2.1. Siamese trackers

In recent years, Siamese networks [10,17,57,64] have received a lot of attention and are widely used in the single object tracking area. One of the most difficult and essential computer vision problems is the accurate and reliable tracking of visual objects [66]. It requires segmentation, or the target's approximate approximation in the form of a bounding box and involves calculating the trajectory of the target in an image sequence given just its initial location. The visual object tracking challenge is approached by Siamese trackers as a matching problem. To achieve tracking during inference, a template from the first frame is clipped and matched to the search regions in the current frame. When used in numerous common tracking

settings, they operate with good performance and real-time tracking speed. However, when the targets experience dramatic changes in appearance, non-rigid deformations, and partial occlusions, their vulnerability becomes apparent. In contrast to previous works, our framework fully utilizes multiple-frame data during the tracking process, which can combine spatial information and temporal information, significantly increasing the model's robustness in those difficult situations.

SiamFC [28] uses the Siamese network as a feature extractor and introduces the correlation layer first in order to combine feature maps. DSiam [72] learns a feature transformation to deal with the target appearance variation and to support the background. RASNet [68] incorporates various attention mechanisms into the Siamese network to adapt the tracking model to the current target. In order to deal with scale variation and get a more accurate target bounding box, The RPN [33] is introduced into the SiamFC by SiamRPN [25]. In diverse methods, SiamRPN++ [29], SiamMask [40] eliminate influence variables like padding from the Siamese network-based visual trackers and incorporate contemporary deep neural networks. Relation Detector (RD) is offered to use a contrastive few-shot learning-based training technique to be able to filter the background distractions [65].ULAST [63] proposes A consistency propagation transformation to generate reliable template kernel.

2.2. DCF trackers

In the recent tracking community, due to their effectiveness and adaptability, discriminative correlation filters (DCF) based approaches have received a lot of attention [2,4,18–20,26,36,41,30,38]. Traditional DCF trackers include MOSSE [18], CSK [19], and KCF [2]. There have been numerous suggestions for improvements to DCF tracking methods, including SAMF [4] and FDSST [20] for scale changes, CN [26] and Staple [36] that account for color information, LCT [26]. The majority of these methods use hand-crafted components, which reduces their robustness and accuracy.

Researchers in the tracking field have begun to concentrate on deep trackers that take advantage of CNN's advantages in object classification [27,31], detection [33], and segmentation [32] tasks, which were inspired by the success of CNN in these tasks. The DCF framework and CNN characteristics are frequently combined because D- CF offers a great framework for contemporary tracking research. In HCF [35] and HDT [37], CNN is used to extract features rather than manually creating them, and the hierarchical response and hedging weak trackers, respectively, are combined to produce the final tracking results. The chosen CNN features are always trained beforehand on various tasks in the methods indicated above, and each component of the tracking systems is learned separately. Thus, the tracking results that were achieved could not be ideal.

2.3. Three-way decision

The three-way decision (3WD) notion was first proposed to explain the three regions of probabilistic rough sets [9]. Three regions—positive region (POS), negative region (NEG), and border region (BND)—are used in Yao's original 3WD proposal. Three-way decision rules are naturally derived from the association of rules derived from the three zones with various actions and decisions [69]. To get a minimal cost ternary classifier, three-way decisions offer a technique to trade off various classification errors [71]. These regions can be regarded as three decision acts, namely acceptance, rejection, and non-commitment. Since 3WD ingeniously offers a delay method, it is consistent with how people think [67]. It has since been expanded into a more general theory by incorporating concepts from interval sets, rough sets, decision-theoretic rough sets, fuzzy sets, and shadowed sets [39]. The concepts of acceptance, rejection and non-commitment serve as the foundation for the 3WD theory. The 3WD eliminates the uncertainty of two-way decision-making by introducing a third alternative and then gathering more data to make a more confident judgment [16,42,58–60]. The studies of 3WD have extended from narrow 3WD to wide 3WD. Numerous generalized models, including three-way approximation models [70]: adopt a generalized definition of three-valued sets by using a set of three values {n, m, p} to replace {0, [0, 1], 1}, three-way analysis models [73], three-way concept lattice models [74]: aims to analyze the uncertainty and incompleteness, which is defined by the truth-membership, indeterminacy-membership, and falsity membership functions of a single-valued neutrosophic set in the given fuzzy attribute set. Recently, based on the prevalent "three" phenomena that exist in the domains of computer sciences, management, cognitive science, and other areas, the wide 3WD has been thoroughly explored. By assessing these probabilities in accordance with the preferences of the user and the structural characteristics of the three partitions, conditional entropy and cross entropy are utilized to choose strategies [75]. Based on the reduction of the conditional entropy, it chooses the course of action or strategy that has the greatest likelihood of producing beneficial motions.

3. Proposed method

In this section, we now introduce our proposed tracking system in detail. In section 3.1, we will introduce the overview of the tracking framework architecture. Then, we will describe each module of the entire framework from section 3.2 to section 3.3.

3.1. Framework architecture

The general structure of our proposed technique, as illustrated in Fig. 2, may be separated into three parts: a historical frame selection sub-network, a deep feature extraction sub-network, and a CF tracking sub-network. The feature extraction

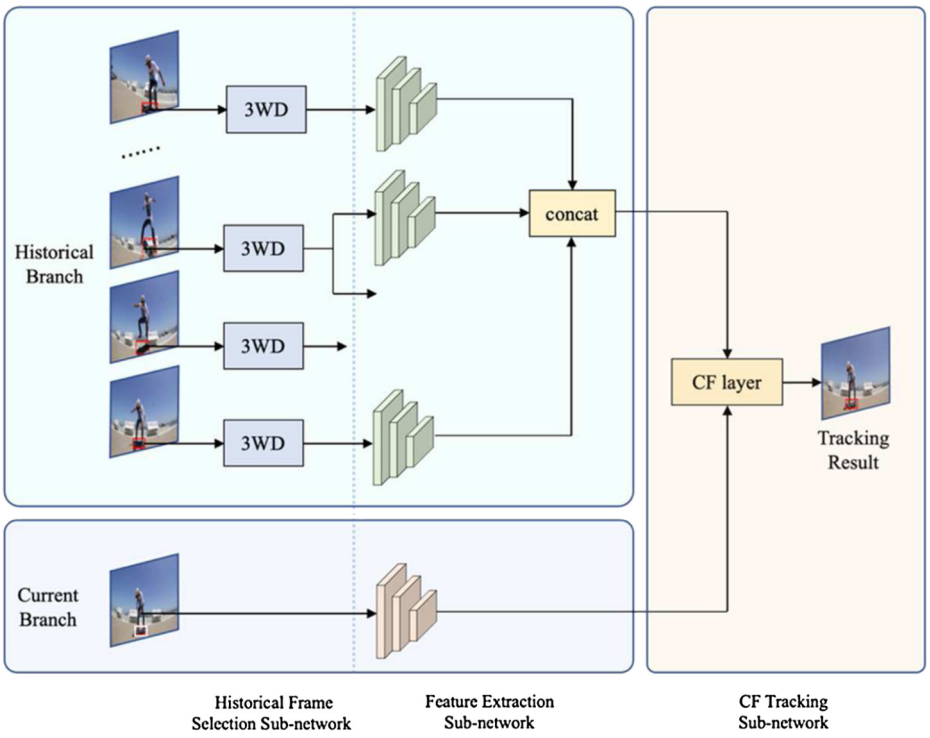

Fig. 2. The overall architecture of our proposed model.

sub-network utilizes Siamese network and could be divided into two parts: the first one is historical branch and the second one is current branch which both use the same backbone to extract deep feature from frames. The difference between the historical branch and the current branch is the input frames and there is an extra historical frame selection sub-network in the historical branch. In the current branch, we set the current frame as the input of the feature extraction sub-network.

The historical frame selection sub-network takes several historical frames as input and utilizes the historical frame selection sub-network which introduces the three-way decision theory to determine whether each historical frame contains enough information to be considered as a useful temporal frame. This operation is enabled because in complex situations, objects become occluded or deformed, and the included features are destroyed. If forced adoption and feature fusion are used, good feature representation will be contaminated, and target tracking will be hampered. After the selection sub-network filter the input frames, we use the backbone network extracts appearance features. The tracking result of historical frame feature maps would then be concat, which means the operation of concatenation along the temporal axis. Finally, after the feature extraction, both the feature maps of the two branches are then sent into the correlation filter layer. A summary of our proposed method is given in Algorithm 1.

Algorithm 1 Spatial-temporal single object tracking with three-way decision theory.

Input: Frames $\{f\}_1^{i+1}$, Tracking results $\{r\}_1^i$
Output: Bounding box prediction r_i
n = 0, p = I
repeat
  judge the f_i by calculating the similarity of r_i, r_q (r_q is the closest frame in $SBS_{(\alpha,\beta)}(S)$)
  if $S(r_i, r_q) \geq \alpha$ **then**
    extract the feature of r_i, n = n+1, p = p-1
  else if $S(r_i, r_q) \leq \beta$ **then**
    j = j+1, p = p-1
  else
    judge the f_i by calculating the similarity of r_i, *template*
  end if
until n = T
concat the selected features $\{F\}_1^T$ to F'
use CF layers to estimate the target bounding box r_i

3.2. Three-way decision based memory frame selection sub-network

When complex situations such as occlusion and fast motions occur, the target object will lose a portion of its feature. Tracking with such samples has a negative impact. As a result, before extracting the feature of historical frames, we introduce the three-way decision theory to ensure that the historical frame contains enough information. First, we compute the similarities S_i between the tracking result of historical frame r_i and the closest tracking result of historical frame r_q in positive domain. We define two parameters α and β as thresholds for dividing strong or weak blocking pairs. If the similarity between the tracking result of historical frame and the closest tracking result of historical frame in positive domain exceed the α, which means the two tracking results are consistent and the r_i consists enough information to help tracking. If the similarity between the tracking result of historical frame and the closest tracking result of historical frame in positive domain below the β, which means the r_i does not consist enough information to help tracking, see (1). (The settings of α and β are described in chapter 4.2)

$$\begin{aligned} SBS_{(\alpha,\beta)}(S) &= \{r_i \in BS \mid S(r_i, r_q) \geq \alpha\}, \\ WBS_{(\alpha,\beta)}(S) &= \{r_i \in BS \mid S(r_i, r_q) \leq \beta\}, \\ BBS_{(\alpha,\beta)}(S) &= \{r_i \in BS \mid \beta < S(r_i, r_q) < \alpha\}. \end{aligned} \tag{1}$$

For each r_i in SBS, we add the historical frame f_i to the category M as one of the memory frames. These frames in category M include sufficient effective target features to aid in locating the target in the current frame. For each r_i in WBS, we add the historical frame f_i to the category N and continue to compute the similarity between r_{i-1} and r_{i-2} until there are T frames in category M. The targets of the frames in category N may suffer from complex situation like occlusion or deformation, and do not contain sufficient information to help to locate the target in the current frame. For each r_i in BBS, we add the historical frame f_i to the category X and compute the similarity of r_i and template to judge if f_i could be utilized as a memory frame. When a frame was set in category X, we need further judgment to conclude whether it have enough information about the target.

3.3. Deep feature extraction sub-network

The deep feature extraction sub-network could be separated into two parts: a historical branch and a current branch. Both the historical branch and the current branch use the same backbone to extract feature.

The inputs of the historical branch are the tracking results of T frames f_i which are selected by the memory frame selection sub-network. Then we adopt ResNet50 φ as the backbone of our proposed method and generate $\{\varphi(r_i), f_i \in M\}$. To enhance the feature representation, we concatenate the feature maps of memory frames to produce the feature map F_{I-1}.

Since it does not require the three-way decision selection step, the current branch differs from the historical branch. Input the research region of the current frame q and utilize the convolutional NETWORK φ *to* generate a corresponding feature map $\varphi(q)$.

The preceding is a detailed introduction to the framework and the specific algorithm of our proposed method. The model consists of three parts: a historical frame selection sub-network, a deep feature extraction sub-network, and a CF tracking sub-network. The feature extraction sub-network could introduce the spatial information since we utilize the deep-learning network as backbone to extract structure features from current frame. The historical frame selection sub-network could introduce the temporal information since the historical frames would be filtered and make full use of.

4. Experiments

Our proposed tracker is implemented in Python using *PyTorch* framework. Experiments are performed on four challenging tracking datasets: OTB-2015 [23], VOT2018 [22], GOT-10k [44] and TrackingNet [43]. All of the tracking results use the reported results to ensure a fair comparison.

4.1. Implementation details

In the training stage, we adopt ImageNet, TrackingNet [43], and COCO [45] as our training dataset and set the input size of search images to 255 $\times$ 255. The pre-trained ImageNet parameters are used to initialize the modified ResNet50, while random parameters are used for the other components. With the mini-batch size of 28, we train the suggested network using 300000 iterations.

4.2. Parameters setting

In the memory frame selection sub-network, to ensure that the historical frame contains enough effective information which can help to locate the target in the current frame, we introduce the three-way decision theory to our tracking

Table 1
Comparisons of our method with different parameter pairs in OTB-2015.

α	β	Success	α	β	Success
0.90	0.60	0.698	0.90	0.70	0.697
0.85	0.60	0.709	0.85	0.70	0.704
0.80	0.60	0.703	0.80	0.70	0.705
0.75	0.60	0.700	0.75	0.70	0.697
0.70	0.60	0.689	0.90	0.75	0.695
0.65	0.60	0.682	0.85	0.75	0.699
0.90	0.65	0.711	0.80	0.75	0.673
0.85	0.65	0.715	0.90	0.80	0.682
0.80	0.65	0.706	0.85	0.80	0.693
0.75	0.65	0.703	0.90	0.85	0.688

Table 2
Comparisons with top trackers published in recent years in OTB-2015. Red, green, and blue indicate the best three tracking results. According on the Success values, trackers are ranked from bottom to top and right to left.

Trackers	Success	Trackers	Success
SiamFC++ [14]	0.682	PrDiMP-50 [6]	0.697
Ocean [54]	0.683	SiamBAN [52]	0.697
DiMP-50 [53]	0.684	SiamCAR [48]	0.698
SPM [50]	0.686	SiamAttn [49]	0.709
PGNet [51]	0.691	RPT [47]	0.712
SaimRPN++ [29]	0.695	Ours	0.715

method. In order to better classify the historical frames and filter out the frames that we need to contain sufficient target information, it is significant to set the appropriate threshold. Therefore, we utilize different thresholds to conduct comparative experiments on dataset OTB-2015. As shown in Table 1, the success of tracking result is best when α is 0.85 and β is 0,65.

4.3. Results on OTB-2015

OTB2013 [46] is a classical benchmark in single object tracking, containing 50 fully annotated sequences that are collected from commonly used tracking sequences. OTB-2015 [23], which has 100 video sequences with 600 frames per video on average, is the extension of OTB-2013. Some of the newer sequences are harder to track. The success plot and precision plot are the two measures used in the evaluation. The precision plot displays the percentage of frames where the tracking results are within a predetermined distance from the ground truth as defined by a set threshold. The representative precision score is always regarded to be the value when the threshold is 20 pixels. When the threshold varies from 0 to 1, the success plot displays the ratios of successful frames. A successful frame is one whose overlap exceeds the specified threshold. Each success plot's area under curve (AUC) is used to rank the tracking method. As shown in Table 2, our proposed approach outperforms all the other trackers in terms of the success (AUC) metric.

4.4. Results on VOT2018

The Visual Object Tracking (VOT) challenges are well-known competitions in single object tracking. They have been held on numerous occasions since 2013, and the outcomes will be presented at ICCV or ECCV. The 2018 version of the visual object tracking challenge contains 60 videos. Following the VOT2018 dataset evaluation protocol, we report our tracker's results in terms of expected average overlap (EAO), accuracy (A), and robustness (R). Table 3 illustrates the accuracy and robustness of our proposed method are worse than some other trackers [54]. However, the expected average overlap (EAO) is similar to SiamBAN [52] and PGNet [51].

4.5. Results on TrackingNet

TrackingNet [43] is a large-scale short-term tracking dataset containing a huge number of videos taken outdoors that are available for training and testing. We evaluate our proposed approach on the testing set which contains 511 videos and obtain results from the dedicated evaluation server. As demonstrated in Table 4, our tracker significantly surpasses all prior state-of-the-art real-time techniques.

Table 3
Comparisons with top trackers published in recent years in VOT2018. Red, green, and blue indicate the best three tracking results. According on the expected average overlap (EAO), trackers are ranked from bottom to top.

Trackers	EAO	A	R
SaimRPN++ [29]	0.412	0.597	0.237
SiamFC++ [14]	0.428	0.583	0.185
ULAST [63]	0.436	0.588	0.168
DiMP-50 [53]	0.440	0.595	0.152
PrDiMP-50 [6]	0.441	0.617	0.168
PGNet [51]	0.447	0.616	0.193
SiamBAN [52]	0.453	0.597	0.178
SiamCAR [48]	0.455	0.598	0.179
SiamAttn [49]	0.472	0.629	0.160
Ocean [54]	0.483	0.594	0.121
Ours	0.450	0.583	0.161

Table 4
Comparisons with top trackers published in recent years in TrackingNet. Red, green, and blue indicate the best three tracking results. According to the "Suc." values, trackers are ranked from bottom to top.

Trackers	Suc.	Prec.	Norm. Prec.
ATOM [55]	69.8	64.3	76.8
SaimRPN++ [29]	72.9	69.1	79.5
DiMP-50 [53]	73.9	68.2	80.0
SiamAttn [49]	75.3	-	81.2
SiamFC++ [14]	75.3	70.6	80.1
PrDiMP-50 [6]	75.9	70.2	81.2
Ours	76.3	71.0	82.3

Table 5
Comparisons with top trackers published in recent years in GOT-10k. Red, green, and blue indicate the best three tracking results. According on the average overlap (AO) and success rates (SR) at thresholds of 0.5 and 0.75, trackers are ranked from bottom to top.

Trackers	AO	$SR_{0.5}$	$SR_{0.75}$
SaimRPN++ [29]	0.518	0.612	0.323
ATOM [55]	0.555	0.633	0.403
SiamCAR [48]	0.569	0.667	0.412
SiamFC++ [14]	0.596	0.693	0.478
DiMP-50 [53]	0.608	0.715	0.495
PrDiMP-50 [6]	0.632	0.732	0.543
Ours	0.640	0.736	0.572

4.6. Results on GOT-10k

A recent large-scale generic object tracking benchmark called GOT-10k [17] contains 10,000 films in total, of which 180 videos make up the testing set. Similar to TrackingNet, the testing set's ground facts are similarly hidden, necessitating the examination of each tracking result in its own evaluation server. GOT- 10k benchmark, in contrast to others, limits trackers to using only the training set for training. This protocol is used in this study to train our tracker and test it on the testing set. The training data is the only change to the settings. In Table 5, we compare our tracker's average overlap (AO) and success rates (SR) at thresholds of 0.5 and 0.75 with those of other trackers.

4.7. Con results on LaSOT

A novel large-scale benchmark, LaSOT [79], for visual tracking which covers 85 object categories and consists of 1550 videos totaling more than 3.87M frames. Each frame is meticulously examined and manually given a bounding box label. Each annotation box is visually examined twice and corrected as necessary to ensure quality. To our knowledge, LaSOT is by far the most comprehensive tracking benchmark with dense annotations (measured in terms of the number of frames). By

Table 6
Comparisons with top trackers published in recent years in LaSOT. Red, green, and blue indicate the best three tracking results. According on the precision (Pre) and success rates (Suc), trackers are ranked from bottom to top.

Trackers	Suc	Pre
SiamCAR [48]	0.363	0.352
SaimFC++ [14]	0.358	0.362
SiamFC [28]	0.336	0.337
SiamRPN [25]	0.409	0.382
ULAST [63]	0.468	0.446
Ours	0.473	0.452

making LaSOT available, we hope to provide the community with a focused platform for the unified training and assessment of tracking algorithms. Long-term tracking evaluation is possible using LaSOT. The average video duration of LaSOT is about 2500 frames, or about 83 seconds, and the shortest and longest sequences, respectively, are 1000 frames and 11,397 frames. This allows for the evaluation of long-term trackers. LaSOT gives visual bounding box annotations and natural language definition, in contrast to current benchmarks that simply provide bounding boxes, which has been demonstrated to be advantageous for a variety of vision tasks, including tracking. In Table 6, we compare our tracker's success rate (Suc) and precision (Pre) with those of other state-of-the-art single object trackers.

5. Conclusions

This work proposes a novel end-to-end single object tracking framework based on spatial-temporal feature combinations. Specifically, our framework abandons the conventional template-based tracking mechanism, and instead locates the target in the current frame by using multiple historical frames. Additionally, we introduce the three-way decision theory to make better historical frame selection. Extensive experiments demonstrate that the proposed method performs better than many other tracking approaches. The effectiveness of our approach is validated in OTB-2015, TrackingNet, GOT-10k, LaSOT and VOT2018 datasets, while running in almost real-time at 22 FPS.

CRediT authorship contribution statement

Ziye Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Duoqian Miao:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgement

This work is supported in part by National Key Research and Development Program No. 2022YFB3104700, the National Science Foundation of China under Grant No. 61976158, the National Science Foundation of China under Grant No. 61976160, the National Science Foundation of China under Grant No. 62076182. This paper is partially supported by the Jiangxi "Double Thousand Plan", and the National Natural Science Foundation of China (Serial No. 62163016), and the Jiangxi Provincial Natural Science Fund (No. 20212ACB202001)}. Appreciate very much for all our reviewers' effort in reading the manuscript in so much detail. We feel very privileged to having you as our reviewer. Thanks!

References

[1] K.H. Lee, J.N. Hwang, On-road pedestrian tracking across multiple driving recorders, IEEE Trans. Multimed. 17 (9) (2015) 1429–1438.
[2] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2014) 583–596.

[3] J. Xing, H. Ai, S. Lao, Multiple human tracking based on multi-view upper-body detection and discriminative learning, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 1698–1701.
[4] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, in: European Conference on Computer Vision, Springer, Cham, 2014, pp. 254–265.
[5] G. Bhat, M. Danelljan, L.V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6182–6191.
[6] M. Danelljan, L.V. Gool, R. Timofte, Probabilistic regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7183–7192.
[7] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4310–4318.
[8] P. Zhang, T. Li, Z. Yuan, X. Yang, Heterogeneous feature selection based on neighborhood combination entropy, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–14.
[9] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: International Conference on Rough Sets and Knowledge Technology, Springer, Berlin, Heidelberg, 2009, pp. 642–649.
[10] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "Siamese" time delay neural network, Adv. Neural Inf. Process. Syst. (1993) 6.
[11] L. Liu, J. Xing, H. Ai, X. Ruan, Hand posture recognition using finger geometric feature, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, 2012, pp. 565–568.
[12] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, SiamCAR: Siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6269–6277.
[13] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware Siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.
[14] Y. Xu, Z. Wang, Z. Li, Y. Yuan, G. Yu, Siamfc++: towards robust and accurate visual tracking with target estimation guidelines, Proc. AAAI Conf. Artif. Intell. 34 (07) (2020) 12549–12556.
[15] Z. Zhang, H. Peng, Deeper and wider Siamese networks for real-time visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4591–4600.
[16] Y. Yang, D. Miao, H. Zhang, 3W-AlignNet: a feature alignment framework for person search with three-way decision theory, Cogn. Comput. (2021) 1–11.
[17] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4353–4361.
[18] D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, Visual object tracking using adaptive correlation filters, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2544–2550.
[19] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: European Conference on Computer Vision, Springer, Berlin, Heidelberg, 2012, pp. 702–715.
[20] M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Discriminative scale space tracking, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2016) 1561–1575.
[21] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, Lasot: a high-quality benchmark for large-scale single object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5374–5383.
[22] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin Zajc, et al., The sixth visual object tracking vot2018 challenge results, in: Proceedings of the European Conference on Computer, Vision (ECCV) Workshops, 2018.
[23] Y. Wu, J. Lim, M.H. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (09) (2015) 1834–1848.
[24] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, et al., The visual object tracking vot2015 challenge results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 1–23.
[25] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with Siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[26] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1090–1097.
[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[28] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H. Torr, Fully-convolutional Siamese networks for object tracking, in: European Conference on Computer Vision, Springer, Cham, 2016, pp. 850–865.
[29] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, Siamrpn++: evolution of Siamese visual tracking with very deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4282–4291.
[30] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, D. Tao, Multi-store tracker (muster): a cognitive psychology inspired approach to object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 749–758.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[32] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[33] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. (2015) 28.
[34] H. Fan, H. Ling, Siamese cascaded region proposal networks for real-time visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7952–7961.
[35] C. Ma, J.B. Huang, X. Yang, M.H. Yang, Hierarchical convolutional features for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3074–3082.
[36] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P.H. Torr, Staple: complementary learners for real-time tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1401–1409.
[37] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M.H. Yang, Hedged deep tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4303–4311.
[38] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1396–1404.
[39] Y. Yao, An Outline of a Theory of Three-Way Decisions. International Conference on Rough Sets and Current Trends in Computing, Springer, Berlin, Heidelberg, 2012, pp. 1–17.
[40] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P.H. Torr, Fast online object tracking and segmentation: a unifying approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.
[41] C. Ma, X. Yang, C. Zhang, M.H. Yang, Long-term correlation tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5388–5396.

[42] W. Shen, Z. Wei, Q. Li, H. Zhang, D. Miao, Three-way decisions based blocking reduction models in hierarchical classification, Inf. Sci. 523 (2020) 63–76.
[43] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, B. Ghanem, Trackingnet: a large-scale dataset and benchmark for object tracking in the wild, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 300–317.
[44] L. Huang, X. Zhao, K. Huang, Got-10k: a large high-diversity benchmark for generic object tracking in the wild, IEEE Trans. Pattern Anal. Mach. Intell. 43 (5) (2019) 1562–1577.
[45] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, in: Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 6–12.
[46] Y. Wu, J. Lim, M.H. Yang, Online object tracking: a benchmark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2411–2418.
[47] Z. Ma, L. Wang, H. Zhang, W. Lu, J. Yin, Rpt: learning point set representation for Siamese visual tracking, in: European Conference on Computer Vision, Springer, Cham, 2020, pp. 653–665.
[48] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, SiamCAR: Siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6269–6277.
[49] Y. Yu, Y. Xiong, W. Huang, M.R. Scott, Deformable Siamese attention networks for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6728–6737.
[50] G. Wang, C. Luo, Z. Xiong, W. Zeng, Spm-tracker: series-parallel matching for real-time visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3643–3652.
[51] B. Liao, C. Wang, Y. Wang, Y. Wang, J. Yin, Pg-net: pixel to global matching network for visual tracking, in: European Conference on Computer Vision, Springer, Cham, 2020, pp. 429–444.
[52] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6668–6677.
[53] G. Bhat, M. Danelljan, L.V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6182–6191.
[54] Z. Zhang, H. Peng, J. Fu, B. Li, W. Hu, Ocean: Object-aware anchor-free tracking, in: European Conference on Computer Vision, Springer, Cham, 2020, pp. 771–787.
[55] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, Atom: accurate tracking by overlap maximization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4660–4669.
[56] Z. Fu, Q. Liu, Z. Fu, Y. Wang, Stmtrack: template-free visual tracking with space-time memory networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13774–13783.
[57] D. Chicco, Siamese neural networks: an overview, J. Artif. Neural Netw. (2021) 73–94.
[58] R. Wan, D. Wang, D. Miao, Gaussian mixture clustering based on three-way decision, J. Chongqing Univ. Post Telecommun. (2021) 806–815.
[59] Z. Wang, D. Miao, C. Zhao, S. Luo, Z. Wei, A robust long-term pedestrian tracking-by-detection algorithm based on three-way decision, in: International Joint Conference on Rough Sets, 2019, pp. 522–533.
[60] G. Lang, D. Miao, M. Cai, Three-way decision approaches to conflict analysis using decision-theoretic rough set theory, Inf. Sci. (2017) 185–207.
[61] Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.
[62] J. Zhang, W. Feng, T. Yuan, J. Wang, A.K. Sangaiah, SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking, Appl. Soft Comput. 118 (2022) 108485.
[63] Q. Shen, L. Qiao, J. Guo, P. Li, X. Li, B. Li, et al., Unsupervised learning of accurate Siamese tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8101–8110.
[64] M. Ondrašovič, P. Tarábek, Siamese visual object tracking: a survey, IEEE Access 9 (2021) 110149–110172.
[65] S. Cheng, B. Zhong, G. Li, X. Liu, Z. Tang, X. Li, J. Wang, Learning to filter: Siamese relation network for robust tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4421–4431.
[66] S. Javed, M. Danelljan, F.S. Khan, M.H. Khan, M. Felsberg, J. Matas, Visual object tracking with discriminative filters and Siamese networks: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–20.
[67] J.H. Zhu, Z.S. Chen, B. Shuai, W. Pedrycz, K.S. Chin, L. Martínez, Failure mode and effect analysis: a three-way decision approach, Eng. Appl. Artif. Intell. 106 (2021) 104505.
[68] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, S. Maybank, Learning attentions: residual attentional Siamese network for high performance online visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4854–4863.
[69] Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (3) (2010) 341–353.
[70] Y. Yao, S. Wang, X. Deng, Constructing shadowed sets and three-way approximations of fuzzy sets, Inf. Sci. 412 (2017) 132–153.
[71] Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (6) (2011) 1080–1096.
[72] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic Siamese network for visual object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1763–1771.
[73] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Resilience analysis of critical infrastructures: a cognitive approach based on granular computing, IEEE Trans. Cybern. 49 (5) (2018) 1835–1848.
[74] P.K. Singh, Three-way fuzzy concept lattice representation using neutrosophic set, Int. J. Mach. Learn. Cybern. 8 (1) (2017) 69–79.
[75] C. Jiang, D. Guo, Y. Duan, Y. Liu, Strategy selection under entropy measures in movement-based three-way decision, Int. J. Approx. Reason. 119 (2020) 280–291.
[76] X. Shen, X. Sui, K. Pan, Y. Tao, Adaptive pedestrian tracking via patch-based features and spatial–temporal similarity measurement, Pattern Recognit. 53 (2016) 163–173.
[77] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, N. Yu, Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4836–4845.
[78] Z. Zhu, W. Wu, W. Zou, J. Yan, End-to-end flow correlation tracking with spatial-temporal attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 548–557.
[79] H. Fan, H. Bai, L. Lin, F. Yang, P. Chu, G. Deng, et al., Lasot: a high-quality large-scale single object tracking benchmark, Int. J. Comput. Vis. 129 (2) (2021) 439–461.
[80] P. Zhang, T. Li, Z. Yuan, C. Luo, G. Wang, J. Liu, S. Du, A data-level fusion model for unsupervised attribute selection in multi-source homogeneous data, Inf. Fusion 80 (2022) 87–103.