A Method of Real-temporal Object Tracking Combined the Temporal Information and Spatial Information

XiaoShuo Jia (gxnujiashuo@163.com)
Guangdong University of Science and Technology https://orcid.org/0000-0002-3091-8918
Zhihui Li
Guangdong University of Science and Technology
Kangshun Li
South China Agricultural University
Shangyou Zeng
Guangxi Normal University

Research Article

Keywords: Single target tracking, Temporal information, Spatial information, LSTM, Triplet Network

Posted Date: April 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-331786/v1

A Method of Real-temporal Object Tracking Combined the Temporal Information and Spatial Information

Xiaoshuo Jia[1], Zhihui Li[1], Kangshun Li[1,2], Shangyou Zeng[3]

[1]*School of Computer Science, Guangdong University of Science and Technology, Dongguan 523079, Guangdong, China*

[2]*College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China*

[3]*School of Electronic Engineering, Guangxi Normal University, Guilin 541004, Guangxi, Guangxi, China*

Abstract*——*The purpose of single target tracking is to accurately and continuously locate a specific object when it is moving. However, when the objects encounter with fast movement, severe occlusion, too small size, and the same local features, the tracking algorithm which based on correlation filter or convolutional neural network will appear the positioning error phenomenon. Aiming at the above problems, this paper designs a single target tracking algorithm: relative temporal spatial network (RTSnet). RTSnet is a multi-thread network that composed of Relative temporal Information Network (RTInet) and Relative Spatial Information Network (RSInet). RTInet is designed on the basis of LSTM, and it has the predictable characteristics of temporal. It mainly obtains the relative temporal information between the frames before and after the target. RSInet, an improved twin network based on the Triplet Network, has the effect of similarity determination which can to obtain the spatial information between the frames before and after the target. In the experiments, the RTSnet is trained by using LASOT data set and verified by using the LASOT test set and The OTB100 data set. In the test set of LASOT, the accuracy of RTSnet reaches 85.5%, StruckSiam reaches 50% and STRCF reaches 54%. Meanwhile, its tracking speed reaches 120fps due to the RTSnet adopts dual-thread operation. On the OTB100 data-set, the accuracy of RTSnet is 81.1%.

***Index Terms*—Single target tracking, Temporal information, Spatial information, LSTM, Triplet Network.**

I. INTRODUCTION

Target tracking is an accurate and real-time positioning process for continuously moving targets. This technology has been widely used in the field of computer vision, such as video surveillance, robots, and drone positioning. According to the number of targets, target tracking can be divided into single target tracking and multiple target tracking. The single target tracking is the most basic theoretical model in target tracking. That model mainly includes tracking algorithms based on correlation filtering (CF) (Henriques J F, et al. 2012; Li Y, Zhu J. 2014; Galoogahi H K, et al. 2017; Wang M, et al. 2017; Possegger H, et al. 2015) and tracking algorithms based on the convolutional neural networks (CNN) algorithms (Ma N, et al. 2018; Sandler M, et al. 2018; Iandola F N, et al. 2016; Szegedy C, et al. 2015; Simonyan K, Zisseman A. 2014). This paper proposes a new single-target tracking algorithm from the perspective of relative information generated in the continuous movement of the target.

The single target tracking algorithm based on CF mainly uses the filter template to perform correlation algorithm processing on each frame of the input image and achieve the continuous positioning effect of the target. Among them, these algorithms include MOSSE (Bolme D S, et al. 2010), KCF (Henriques J F, et al. 2015), SRDCF (Danelljan M, et al. 2015), DSST (Danelljan M, et al. 2014), etc. These algorithms have certain limitations in obtaining image features. When the background of the moving target is too complex, the accuracy of image feature acquisition by this type of algorithm will be greatly reduced. Since 2012, experts in this field have proposed many single target tracking algorithms based on CNN, such as SiamFC (Bertinetto L, et al. 2016), CFNet (Valmadre J, et al. 2017), SiamRPN (Li B, et al. 2018), GOTURN (Held D, et al. 2016), etc. Compared to tracking algorithms based on CF, the single target tracking algorithms based on CNN has achieved good results in the key indicators of tracking speed and accuracy. The single target tracking algorithms based on CNN has been greatly improved in accuracy and speed, but when the target encounter with fast

motion, severe occlusion, and the same local feature, the accuracy of this type of algorithm will drop quickly. At the same time, the single target tracking algorithms based on CF cannot achieve real-time results in accuracy and speed.

In response to the phenomenon mentioned above, this paper designs a relative temporal spatial information network (RTSnet) based on the relative spatial information and the relative temporal information of the consecutive moving target. RTSnet is a combination of the relative spatial network (RSInet) and the relative temporal network (RTInet). Among RTSnet, RSInet, a trilinear network designed based on Triplet Network (Hoffer E, Ailon N. 2015), mainly make use of the relative spatial information among the targets of three consecutive to predict the spatial information of the next target. RTInet is a time prediction network based on LSTM algorithm (Yang D D, et al. 2016; Zhou J, Xu W. 2015; Greve R, et al. 2016; Gulcehre C, et al. 2016), which can predict the temporal information of the next target through the relative temporal information among the targets of three consecutive. Both RTInet and RSInet make predictions about the target information of the next frame. Therefore, in order to improve the efficiency of the overall algorithm, RTInet and RSInet will do multi-thread parallel operations. Then, RTSnet fuse the spatial information and the temporal information about the target from the RSInet and the RTInet respectively to obtain the target information about the next frame.

The main contributions of this paper can be summarized as follows:

1. This paper proposes a tracking algorithm that fuses spatial information and temporal information, which can obtain more information about moving targets to improve the efficiency of tracking.
2. In the extraction of spatial information, this paper uses the tri-linear twin network to improve the accuracy of similarity determination.
3. In the extraction of temporal information, this paper uses the LSTM algorithm to achieve the prediction characteristics of continuous moving objects and improve the tracking speed.
4. Experiments on LASOT (Fan H, et al. 2018) and OTB100 (Wu Y, et al. 2015) data sets prove that spatial information is beneficial to target tracking. At the same time, when comparing with some state of the art trackers algorithms, it demonstrates the superiority of the algorithm in this paper.

II. Dataset and Prior-work

This section mainly introduces the relevant data sets and the design ideas of RTInet and RSInet, and explains the network structure of RTSnet.

A. Dataset

LASOT is a long-term tracking data set. The data set has 1400 video sequences and each video has an average of 2512 frames. The shortest video in this data set also has 1000 frames, and the longest contains 11397 frames.

The OTB100 data set contains 98 videos and 100 test scenarios. The coordinates of the object to be located in the picture are recorded in the groundtruth_rect of the data set. Each line of coordinates corresponds to the coordinate position of the upper left corner of the positioning frame and the width and height of the positioning frame in the picture.

TD: We first extract the specific location information about the target from both LASOT and OTB100. Then we subtract the current location information from the location information of the next moment and obtain the relative time information about the target. The collection of these relative time information is the data set TD.

SD: We first crop the images of each sequence in LASOT and OTB100 according to the target position at the corresponding time, and the image set obtained after the interception is the spatial information data set of the target.

B. RTInet

We can draw a conclusion from literature (Zhang Y, et al. 2018) that the position information of moving object in temporal has certain continuous features. We will further calculate the relative temporal information of the target on the basis of that conclusion. As shown in FIGURE 1, a is the target position sequence of airplane-6 from the LASOT and b is the relative position sequence of the target. It can be seen from the FIGURE 1 that the data fluctuation of a presents discrete features, while the b presents a relatively stable effect. Therefore, we can obtain more accurate the relative temporal feature information about the target through the calculation method of b.

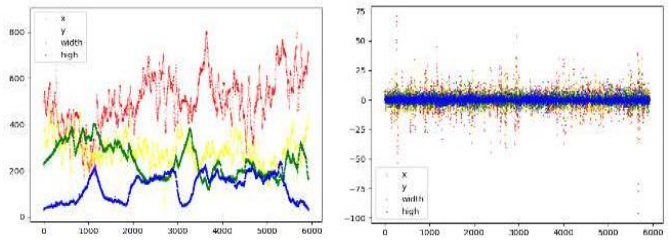

FIGURE 1 TWO DISTRIBUTION METHODS OF SEQUENCE AIRPLANE-6.

On the basis of the above conclusion, we design an RTInet algorithm which can extract the relative temporal information, as depicted in FIGURE 2.

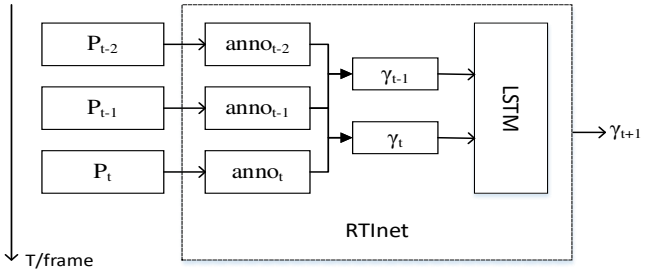

FIGURE 2 RTInet ALGORITHM STRUCTURE DIAGRAM

RTInet is designed on the basis of the LSTM algorithm and trained on the TD dataset. The data in the TD dataset is obtained by $anno_t$ subtract $anno_{t-1}$, which $anno_t$ and $anno_{t-1}$ are the temporal information about the target at the current moment and the previous moment respectively, as shown in formula 1. We will get the $\vec{r}_t|(x,y)$ by the center point (x, y) of the target position from each frame and do the subtraction between the current frame and the previous frame. The data of $\vec{r}_t|(x,y)$ indicates in which direction the target position of the next frame will be located. Then, we put the absolute value of r_t and r_{t-1} into RTInet to calculate the predicted value r_{t+1} of the next frame. r_{t+1} is an absolute value data. When r_{t+1} is added to $anno_t$, we will get prediction data about the next frame in 16 different directions. In order to obtain more accurate prediction data, we will further to filter these data according to the angle of $\vec{r}_t|(x,y)$ and the relative spatial information.

$$r_t = anno_t - anno_{t-1} \tag{1}$$

$$\vec{r}_t|(x,y) = \overrightarrow{anno_t}|(x,y) - \overrightarrow{anno_{t-1}}|(x,y) \tag{2}$$

$$r_{t+1} = RTInet(r_{t-1}, r_t) \tag{3}$$

RTInet first obtains the target's temporal information from three consecutive frames P_{t-2}, P_{t-1}, P_t, namely the target's position information $anno_{t-2}$, $anno_{t-1}$ and $anno_t$. Then, the relative temporal information r_{t-1}, r_t is obtained by formula 1. Finally, the relative temporal information r_{t-1} and r_t are calculated by rti to obtain the relative temporal information r_{t+1} of the next frame. RTInet mainly predicts the relative temporal information of the target in the next frame from the relative temporal information of three consecutive frames.

C. *RSInet*

The moving target not only has temporal information, but also has spatial information. For example, SiamFC and SiamRFC use the spatial information of the target to achieve continuous positioning of the target. When we analyze the moving target, we will divide each image into foreground and background to analyze separately. The foreground mainly describes the spatial information of the target, and the background mainly describes the background information of the target. If the spatial information about the target is obtained from the entire image, it will be disturbed by the background information so that will not only affect the speed of calculation, but also affect the accuracy of the target information acquisition. As for these problems, we propose relative spatial information and design a relative spatial information network, RSInet.

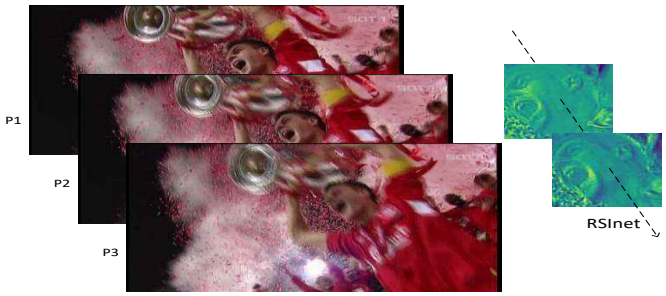

FIGURE 3 RELATIVE SPATIAL INFORMATION OF TARGETS IN CONTINUOUS SEQUENCE

As shown in FIGURE 3, we first crop three corresponding target images from three consecutive moving images, so that we can only focus on the spatial information of the target and avoid the interference of background information. After the three intercepted target images are subtracted between the front and rear frames, two relative images about the target will be obtained. Here, RSInet is used to learn the relationship between consecutive relative images to predict the information of the next frame of target image and the RSInet's process is shown as FIGURE 4.

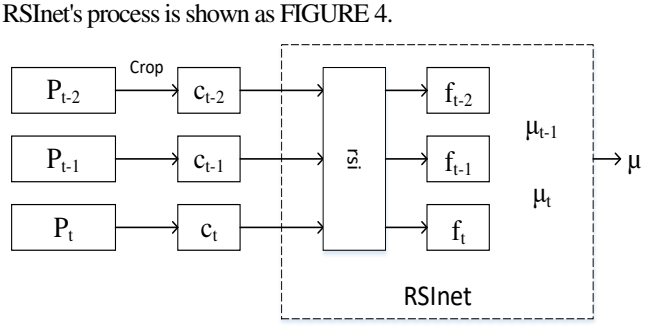

FIGURE 4 RSInet ALGORITHM STRUCTURE DIAGRAM

RSInet firstly extracts three target images information c_{t-2}, c_{t-1}, c_t

from the three consecutive frames of images P_{t-2}, P_{t-1}, P_t according to the target position. Then we put c_{t-2}, c_{t-1}, c_t into rsi which is the feature extraction part of RSInet. And rsi is mainly a feature extraction layer composed of 3*1 and 1*3 types of convolutional layers. The parameters are shown in the TABLE 1. Images c_{t-2}, c_{t-1}, c_t are extracted by rsi to obtain corresponding high-dimensional features f_{t-2}, f_{t-1}, f_t. The high-dimensional feature f_{t-2}, f_{t-1}, f_t obtain the parameter μ in the following formula 4. The parameter μ mainly represents the relative information ratio of the spatial feature information generated by the target between three consecutive frames. We can get the relative vector $\overrightarrow{r_{t+1}}|(x,y)$ of the target center position in the next frame by multiplying $\vec{r}_t|(x,y)$ and μ.

$$\mu = \mu_t / \mu_{t-1} \tag{4}$$

RSInet mainly obtains the spatial relative information of the target from three consecutive frames and predicts the target information in the next frame.

TABLE 1 CONVOLUTION PARAMETERS OF rsi IN RSInet

-	*Kernel size/ stride pad/number*	-	-	*Kernel size/ stride /pad*	-
Net1	*3*3/2/0/32*	*Concat*	*Avg_pool(2*2/1/0)*	*3*3/1/100*	*Avg_pool(3*3/2/0)*
	*1*3/1/0/64*				
	*3*1/1/0/64*				
	*3*3/2/0/32*				
Net2	*3*3/2/0/16*				
	*3*3/1/0/16*				
	*3*3/2/0/32*				
Net3	*3*3/2/0/32*				
	*1*3/1/0/64*				
	*3*1/1/0/64*				
	*3*3/2/0/32*				

D. RTSnet

RTInet extracts the relative temporal information about the target from three consecutive frames and predicts the relative temporal information about the target in the next frame. RSInet extracts relative spatial information about the target from three consecutive frames and predicts the relative spatial information about the target in the next frame under rsi. Here we use multi-threaded parallel computing to fuse the two types of information, as show in FIGURE 5.

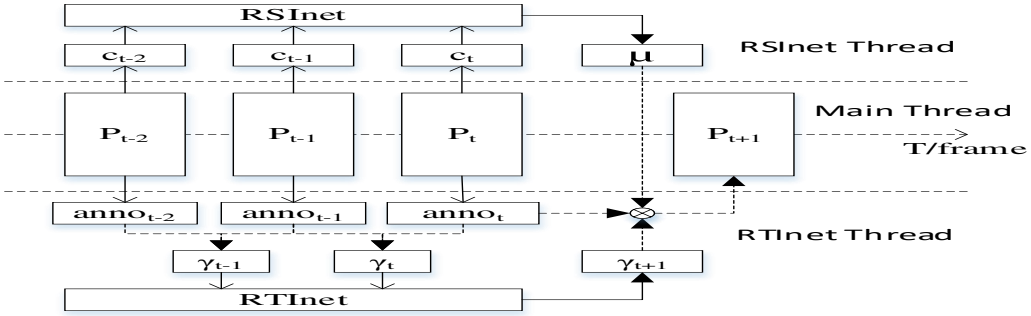

FIGURE 5 RTSnet ALGORITHM STRUCTURE DIAGRAM

The main thread mainly reads information about the image. The RSInet thread reads the target's spatial information c_{t-2}, c_{t-1}, c_t from P_{t-2}, P_{t-1}, P_t and obtains the relative spatial information ratio μ of the target in the next frame. The RTInet thread reads the relative temporal information r_{t-1}, r_t about the target from P_{t-2}, P_{t-1}, P_t and obtains the relative temporal information r_{t+1} about the target in the next frame. Then, to calculate the target information $anno_t$ at the current moment, the relative temporal information r_{t+1}, and the relative spatial information ratio μ through the following formula to obtain the position information $anno_{t+1}$ about the target in the next frame.

III. Experiments

In this section, we discuss the details of RTSnet in training. And compared with some state of the art algorithms from two perspectives of algorithm accuracy and tracking speed under LASOT test set and OTB100 test set respectively.

A. *Experiments preparation*

Experimental environment: The test platform of the network is AMD Ryzen 3 2200G; the training platform of the network is GeForce RTX 1070.

Training details: The purpose of RSInet is to obtain the relative spatial information of the next frame through the relative spatial information between successive frames. When we train RSInet, we first extract the target image P_{t-1} and P_t of two consecutive frames from each sequence and designate them as similar images, then extract the target image P_{t+50} of the 50th frame after P_t as non-similar images. Finally, we train RSInet by making the distance between P_t and P_{t+50} much greater than the distance between P_t and P_{t-1}. When the value range of μ is between (0.25, 0.75), the accuracy of image similarity determination is the highest by experiment. RTInet generates relative temporal information about the next frame through the relative temporal information between successive frames. When we train RTInet, we first extract the temporal information of four consecutive frames from the sequence, namely the location information of the target, $anno_{t-2}$, $anno_{t-1}$, $anno_t$, $anno_{t+1}$. Then, the relative temporal information of γ_{t-1} 、 γ_t and γ_{t+1} between successive frames was obtained. Finally, RTInet was trained in the way of γ_t and γ_{t-1} fitting prediction γ_{t+1}.

Training: RTSnet's algorithm is composed of RSInet and RTInet, and RTSnet's algorithm also will be divided into two parts for training. RTInet's algorithm is trained under TD dataset and optimized under formula 5, aiming to make the predicted value of RTInet close to the true value.

$$L_{rti} = \sqrt{\sum_{i=4}^{4} RTInet(r_{t-1}, r_r)_i^2 - r_{t+1_i}^2} \,/\, 4 \tag{5}$$

RSInet's algorithm is trained under SD dataset and optimized under formula 6, aiming to let RSInet's algorithm to reduce the relative distance $\Delta f_{t-1,t-2}$ between similar image pairs and enlarge the relative distance $\Delta f_{t,t-1}$ between non-similar image pairs.

$$L_{rsi} = \Delta f_{t-1,t-2} - 2 * \Delta f_{t,t-1} \tag{6}$$

B. *Experimental results on LASOT*

We utilize the LASOT toolkit to evaluate the tracking effect of the RTSnet and its contrasting algorithm. The evaluation criteria mainly depend on two aspects: precision plots and success plots. The success plots is that under a threshold, we firstly calculate the IOU (Ren S, et al. 2017) overlap between the predicted border of each frame and the ground truth of the frame, and then compare the comparison rate between the IOU overlap and the threshold. The the precision plots is that the ratios of frames where the location errors are within a certain values. We sort the evaluated trackers according to the area size under the curve scores of the success graph and the accuracy graph.

In order to verify the efficiency and accuracy of RTSnet, we compare the effect of RTSnet with some state of the art tracking networks, including MDNet (Nam H, Han B. 2016), Staple (Bertinetto L, et al. 2016), SRDCF, CSRDCF (Lukezic A, et al. 2017), Struck (Hare S, et al. 2015), KCF, ECO (Daneljan M, et al. 2017), VITAL (Song Y, et al. 2018), StruckSiam (Zhang Y, et al. 2018), D-Siam (Guo Q, et al .2017), SiamFC, etc. Among them, SRCDF, CSRDCF, and KCF are all tracking algorithms based on CF. StruckSiam, D-Siam, SiamFC all use the characteristics of Siamese algorithm to achieve efficient tracking effects. The experimental results are shown in TABLE 2.

TABLE 2 THE ACCURACY OF RTSnet AND SOME STATE OF THE ART TRACKING NETWORKS UNDER THE LASOT DATASET

Network	AUC(%)
RTSnet	**0.855**
Struck-Siam	0.356
D-Siam	0.353
CSR-DCF	0.263
SRDCF	0.271
Struck	0.243

The FIGURE 6 shows the success plots on LASOT dataset. It can be seen intuitively from the figure 6 that the proposed RTSnet has an accuracy rate of 85.5%. And, the proposed RTSnet is 50% and 54% higher than StruckSiam and STRCF respectively.

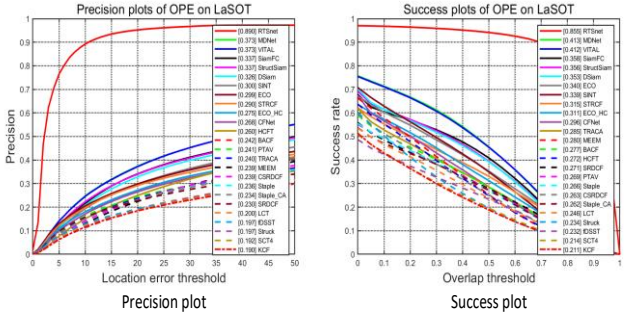

FIGURE 6 THE COMPARISON RESULTS OF RTSnet AND SOME STATE OF THE ART TRACKING NETWORKS UNDER THE LASOT TOOLKIT.

C. *Experimental results on OTB100*

The OTB100 data set contains 98 video sequences and 11 attributes. Here we use the OTB100 toolkit to compare the RTSnet and SiamFC, CFNet, CSRDCF, SRDCF and KCF to find which algorithm has the best accuracy. The result is shown in the FIGURE 7 and TABLE 3. It can be concluded from the comparison result graph that RTSnet has an accuracy rate of 81.1% higher than other tracking algorithms, 21% higher than SiamFC and 19% higher than CFNet.

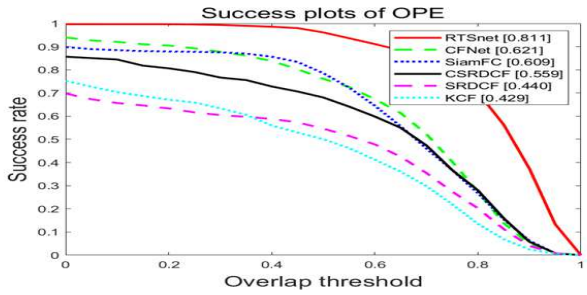

FIGURE 7 RTSnet'S SUCCESS RATE GRAPH UNDER THE OTB100 DATA SET

TABLE 3 THE ACCURACY OF RTSnet AND SOME STATE OF THE ART TRACKING NETWORKS IN THE OTB100 DATASET

Networks	AUG(%)
RTSnet	**0.811**
CFNet	0.621
CSR-DCF	0.569
SRDCF	0.440
SiamFC	0.609

CSRDCF, SRDCF and KCF are all tracking algorithms based on CF, which are used as templates. In terms of feature extraction capability of target spatial information, correlation filter algorithm is lower than CNN algorithm, but in terms of computing speed, it is superior to CNN algorithm, and its hardware requirements are lower than CNN algorithm. Therefore, the tracking algorithm based on correlation filter can be applied to the design of mobile hardware.

IV. Discussion

In the previous section, we discuss the advantages of RTSnet numerically. In this section, we compare RTSnet with some state of the art algorithms under the 14 attributes of LASOT dataset, and make a detailed discussion on the actual tracking effect.

A. *Attribute evaluation*

All sequences in the LASOT dataset contain a total of 14 different attributes, such as illumination variation(IV), full occlusion(FOC), partial occlusion(POC), deformation(DEF), motion blur(MB), fast motion(FM), scale variation(SV), camera motion(CM), rotation(ROT), background clutter(BC), low resolution(LR), viewpoint change(VC), out-of-view(OV), aspect ratio change(ARC).

Here we first analyze the superiority of the data set attributes. Here we use the accuracy in FIGURE 6 as the benchmark, and then calculate the relative changes in the accuracy of the three tracking algorithms RTSnet, StruckSiam and STRCF under 14 attributes as show in FIGURE 8.

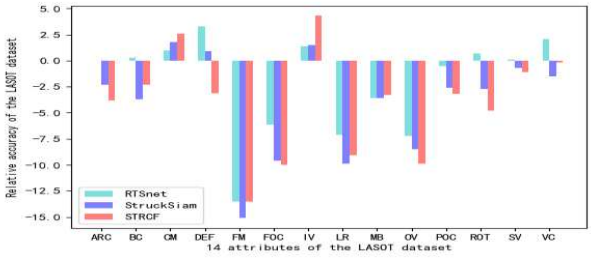

FIGURE 8 UNDER THE 14 ATTRIBUTES OF THE LASOT DATA SET, THE RELATIVE ACCURACY COMPARISON CHART OF RTSnet, STRUCKSIAM AND STRCF

From the accuracy changes in FIGURE 8, we find that under the attributes of FM, FOC, LR, and OV, the three algorithms all have achieved a relatively large gap, but the change value of RTSnet is lower than the other two algorithms. That is mainly because of RTSnet not only adds relative spatial information in the tracking process, but also adds relative temporal information, while the StruckSiam and STRCF algorithms only add spatial information in the tracking process. In the comprehensiveness of information, RTSnet is superior to the other two algorithms. Under the attributes of ARC, BC, ROT, VC, and SV, RTSnet has a positive change, while StruckSiam and STRCF both have a negative change. This is mainly because of BC, ARC, ROT, SV, VC are mainly related to unfavorable factors such as background interference and viewing angle changes. RTSnet can exploit RTInet to obtain temporal feature information, and obtain spatial features about target by RSInet. So RTSnet can reduce the interference of background, perspective and other factors. StruckSiam is based on the principle of CNN to obtain high-dimensional features of the image, but under background interference and perspective changes, CNN will have high errors in extracting target features. However, the STRCF based on the correlation filter will use the spatial

information of the preceding and following frames to locate the target, thus reducing a certain error.

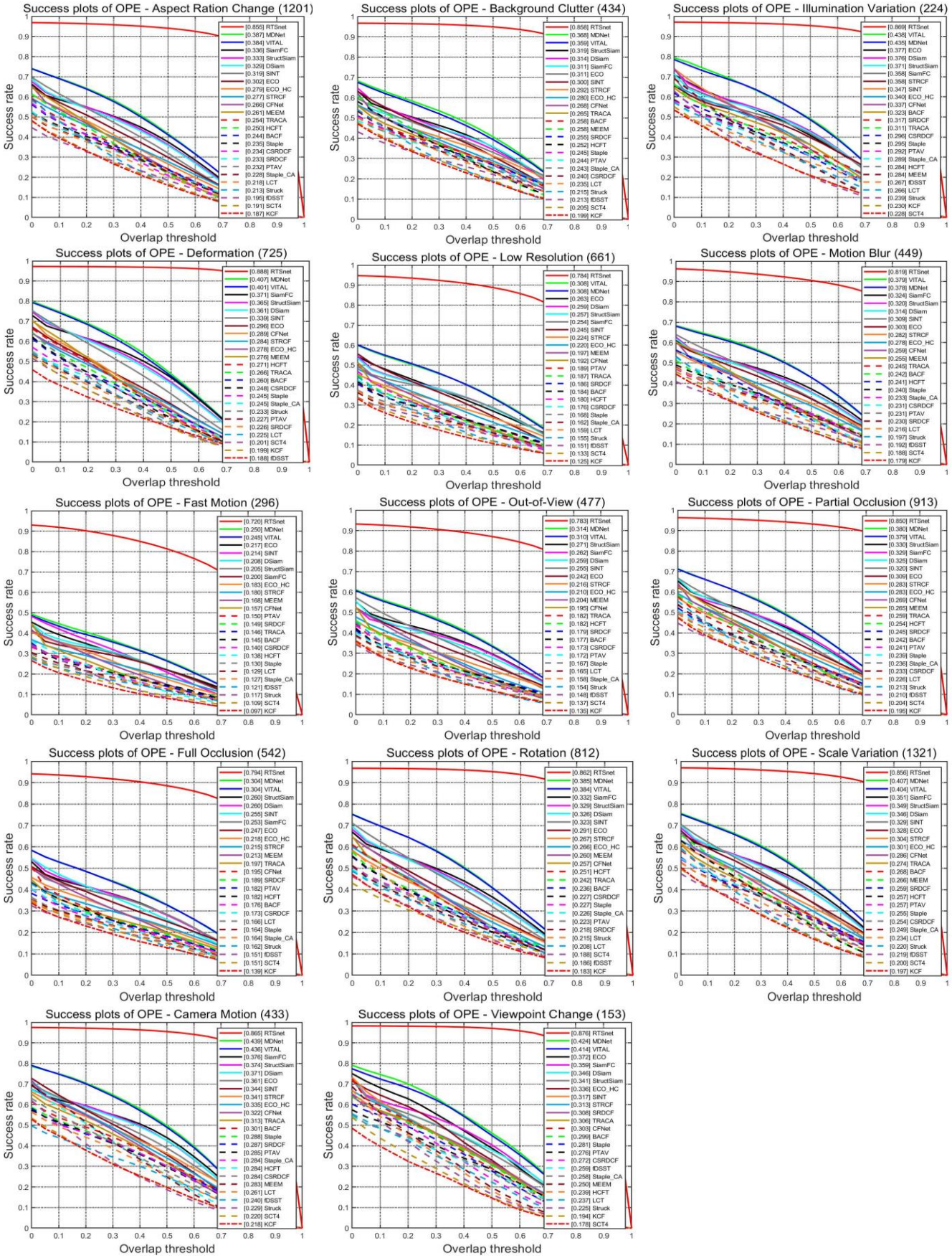

FIGURE 9 UNDER THE 14 ATTRIBUTES OF LASOT, THE ACCURACY COMPARISON CHART OF RTSNET AND SOME ALGORITHMS

Under the MB attribute, the accuracy of the three algorithms is reduced due to the movement of the lens. Although the accuracy of the three algorithms has been lowered under the POC attribute, RTSnet can still perform high-accuracy real-temporal tracking mainly, that's because of the temporal feature information of the target is added to RTSnet. So, RTSnet can reduce the influence of some occlusion factors. On the contrary, the accuracy of StruckSiam and STRCF are drop greatly by the occlusion.

When encountering OV, FM, LR, FOC, StruckSiam and STRCF have serious shortcomings in obtaining target information. On the contrary, RTSnet with the help of RTInet not only to obtain more temporal information about the target, but also to make up for the problem of insufficient spatial information acquisition.

Then, we continue to analyze the success plots of RTSnet and the comparison network under the LASOT data set which are shown in the FIGURE 9. It can be intuitively concluded from the FIGURE 9 that RTSnet has achieved higher accuracy in all attributes. The main reason is that compared with traditional algorithms, RTSnet performs multiple types of information fusion in information acquisition.

B. Qualitative evaluation

Here's a further discussion of the tracking effects of RSTnet. We selected 4 data sets of video sequences from LASOT, namely Basketball-6, Bear-6, Boat-6, Bottle-6.

FIGURE 10 IN 4 DIFFERENT SEQUENCES, THE TRACKING EFFECT COMPARISON CHART OF RTSnet, STRUCKSIAM AND STRCF

As shown in the FIGURE 10, the RTSnet algorithm is compared with the StrcukSiam and STRCF algorithms in the tracking effect. In the Basketball-6 video sequence, it is intuitively found that SturckSiam has severe target loss due to long-term target tracking. In the case of high-speed ball movement and complete occlusion, STRCF also has short-term target loss. Compared with Groundtruth, RTSnet can achieve short-term prediction and tracking of the ball's motion, and is not affected by obstructions and fast motivation. In the Bear-6 video sequence, StuckSiam will cause target tracking errors due to too many targets with the same characteristic. RTSnet and STRCF still can track the target in real temporal with the case of camera shake. In the Boat-6 video sequence, we can intuitively see that because of the target size is too small, StruckSiam and STRCF cannot achieve precise positioning of the target. On the contrary, RTSnet can realize real-temporal tracking on small objects, which based on the characteristic relationship between the front and rear frames. There are 4 identical bottles in the video sequence of Bottle-6. When one bottle is falling down and is partially occluded by another bottle, SturckSiam and STRCF may have target positioning errors due to the similar

characteristics of the bottle image. On the contrary, RTSnet can achieve the tracking effect of the bottle according to the change of the spatial feature information of the target's previous and subsequent frames and the change of the temporal feature information between the previous and subsequent frames.

V. Conclusion

In this paper, we propose a single-target tracking network RTSnet that integrates temporal information and spatial information. RTSnet is composed of the spatial relative network RSInet and the temporal relative network RTInet. First, we analyze the relative information changes of continuous moving targets from the spatial and temporal perspectives. After that, the relative spatial information is obtained from RSInet, and the relative temporal information is obtained from RTInet. Finally, the two kinds of relative information are merged and the prediction information about the next frame of the target is obtained. Our algorithm effectively improves the accuracy and robustness of tracking by acquiring temporal and spatial information of moving targets. Therefore, the proposed tracker is higher robust to DEF, FM, FOC, and SV. As a result, the performance of RTSnet tracker in the LASOT data set and OTB100 data set is superior to some state of the art trackers. Meanwhile, RTSnet uses dual-threaded operation so that the tracking speed on the LASOT data set reaches 120fps.

Acknowledgments

This work is supported by The Natural Science Foundation of Guangdong Province under the Grant No.2020A1515010784.

Author Contributions Xiaoshuo Jia conceived the algorithms, conducted experimental demonstrations, and wrote the paper; **Zhihui Li** wrote the paper; **Kangshun Li** wrote the paper; **Shangyou Zeng** wrote the paper.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Henriques J F, Caseiro R, Martins P, et al (2012) Exploiting the circulant structure of tracking-by-detection with kernels[C]. european conference on computer vision, pp 702-715.
2. Li Y, Zhu J (2014) A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration[C]. european conference on computer vision, pp 254-265.
3. Galoogahi H K, Fagg A, Lucey S, et al (2017) Learning Background-Aware Correlation Filters for Visual Tracking[C]. International conference on computer vision, pp 1144-1152.
4. Wang M, Liu Y, Huang Z, et al (2017) Large Margin Object Tracking with Circulant Feature Maps[C]. Computer vision and pattern recognition, pp 4800-4808.
5. Possegger H, Mauthner T, Bischof H, et al (2015) In defense of color-based model-free tracking[C]. Computer vision and pattern recognition, pp 2113-2120.
6. Ma N, Zhang X, Zheng H T, et al (2018) ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design [J]..
7. Sandler M, Howard A, Zhu M, et al (2018) Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation [J].
8. Iandola F N, Han S, Moskewicz M W, et al (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [J].
9. Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions[C]. Computer vision and pattern recognition, pp 1-9.
10. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition [J].
11. Bolme D S, Beveridge J R, Draper B A, et al (2010) Visual object tracking using adaptive correlation filters[C]. Computer vision and pattern recognition, pp 2544-2550.
12. Henriques J F, Caseiro R, Martins P, et al (2015) High-Speed Tracking with Kernelized Correlation Filters [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. pp 37(3): 583-596.
13. Danelljan M, Hager G, Khan F S, et al (2015) Learning Spatially Regularized Correlation Filters for Visual Tracking[C]. International conference on computer vision, pp 4310-4318.
14. Danelljan M, Hager G, Khan F S, et al (2014) Accurate Scale

Estimation for Robust Visual Tracking [C]. british machine vision conference.
15. Bertinetto L, Valmadre J, Henriques J F, et al (2016) Fully-Convolutional Siamese Networks for Object Tracking[C]. European conference on computer vision, pp 850-865.
16. Valmadre J, Bertinetto L, Henriques J F, et al (2017) End-to-End Representation Learning for Correlation Filter Based Tracking[C]. Computer vision and pattern recognition, pp 5000-5008.
17. Li B, Yan J, Wu W, et al (2018) High Performance Visual Tracking with Siamese Region Proposal Network[C]. Computer vision and pattern recognition, pp 8971-8980.
18. Held D, Thrun S, Savarese S, et al (2016) Learning to Track at 100 FPS with Deep Regression Networks[C]. european conference on computer vision, pp 749-765.
19. Hoffer E, Ailon N (2015) Deep Metric Learning Using Triplet Network [J].
20. Yang D D, Cai Y Z, Mao N, et al (2016) Long-term object tracking based on kernelized correlation filters [J]. Optics and Precision Engineering, pp 24(8):2037-2049.
21. Zhou J, Xu W (2015) End-to-end learning of semantic role labeling using recurrent neural networks[C]. International joint conference on natural language processing, pp 1127-1137.
22. Greve R, Jacobsen E J, Risi S, et al (2016) Evolving Neural Turing Machines for Reward-based Learning[C]. Genetic and evolutionary computation conference, pp 117-124.
23. Gulcehre C, Chandar S, Cho K, et al (2016) Dynamic Neural Turing Machine with Soft and Hard Addressing Schemes.[J].
24. Fan H, Lin L, Yang F, et al (2018) LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking [J]. arXiv: Computer Vision and Pattern Recognition.
25. Wu Y, Lim J, Yang M H (2015) Object Tracking Benchmark [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp 37(9):1834-1848.
26. Zhang Y, Wang L, Qi J, et al (2018) Structured Siamese Network for Real-temporal Visual Tracking[C]. European conference on computer vision, pp 355-370.
27. Ren S, He K, Girshick R B, et al (2017) Faster R-CNN: Towards Real-temporal Object Detection with Region Proposal Networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 39(6): 1137-1149.
28. Nam H, Han B (2016) Learning Multi-domain Convolutional Neural Networks for Visual Tracking[C]. Computer vision and pattern recognition, pp 4293-4302.
29. Bertinetto L, Valmadre J, Golodetz S, et al (2016) Staple: Complementary Learners for Real-temporal Tracking[C]. Computer vision and pattern recognition, pp 1401-1409.
30. Lukezic A, Vojir T, Zajc L C, et al (2017) Discriminative Correlation Filter with Channel and Spatial Reliability[C]. Computer vision and pattern recognition, pp 4847-4856.
31. Hare S, Golodetz S, Saffari A, et al (2015) Struck: Structured Output Tracking with Kernels [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 2096-2109.
32. Danelljan M, Bhat G, Khan F S, et al (2017) ECO: Efficient Convolution Operators for Tracking[C]. Computer vision and pattern recognition, pp 6931-6939.
33. Song Y, Ma C, Wu X, et al (2018) VITAL: VIsual Tracking via Adversarial Learning[C]. Computer vision and pattern recognition, pp 8990-8999.
34. Zhang Y, Wang L, Qi J, et al (2018) Structured Siamese Network for Real-temporal Visual Tracking[C]. European conference on computer vision, pp 355-370.
35. Guo Q, Feng W, Zhou C, et al (2017) Learning Dynamic Siamese Network for Visual Object Tracking[C]. International conference on computer vision, pp 1781-1789.

Figures

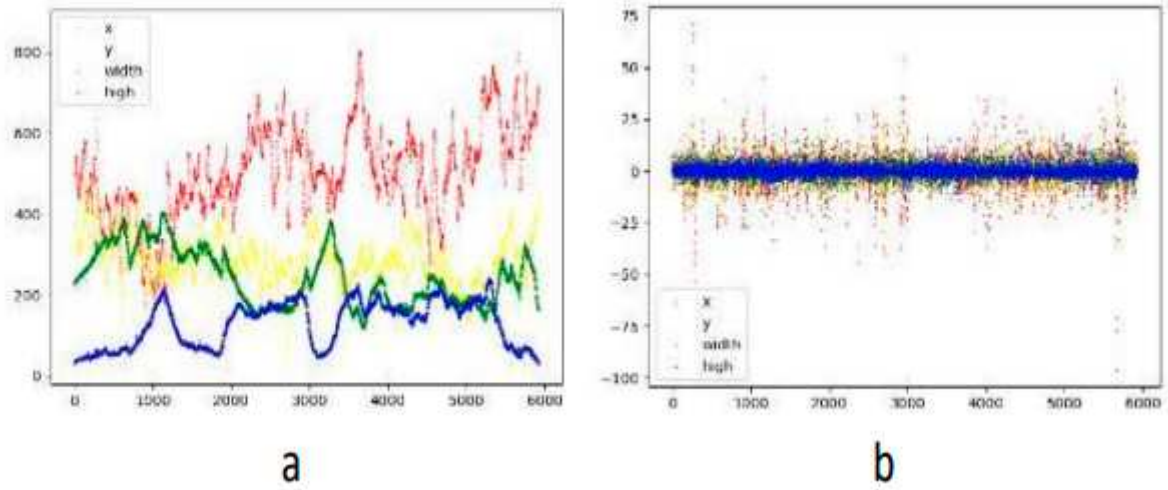

Figure 1

TWO DISTRIBUTION METHODS OF SEQUENCE AIRPLANE-6.

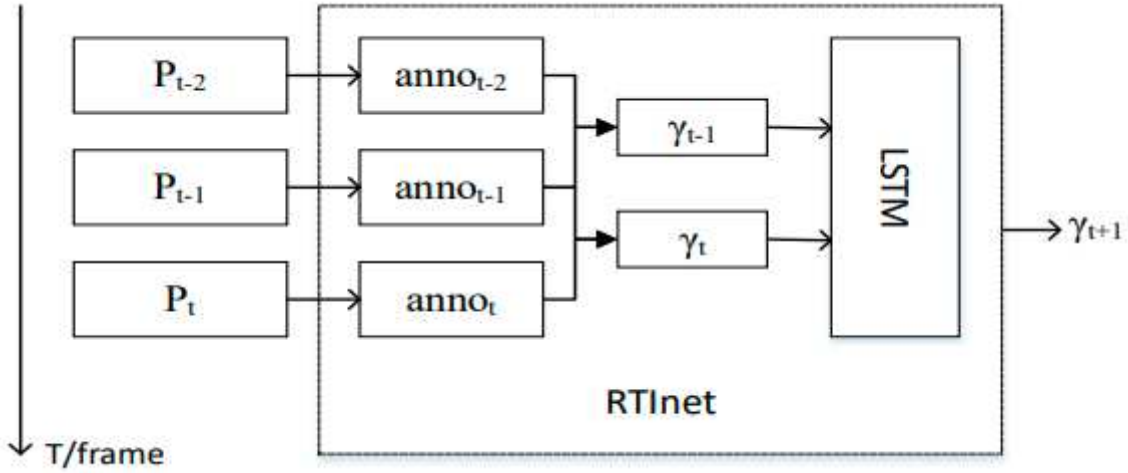

Figure 2

RTInet ALGORITHM STRUCTURE DIAGRAM

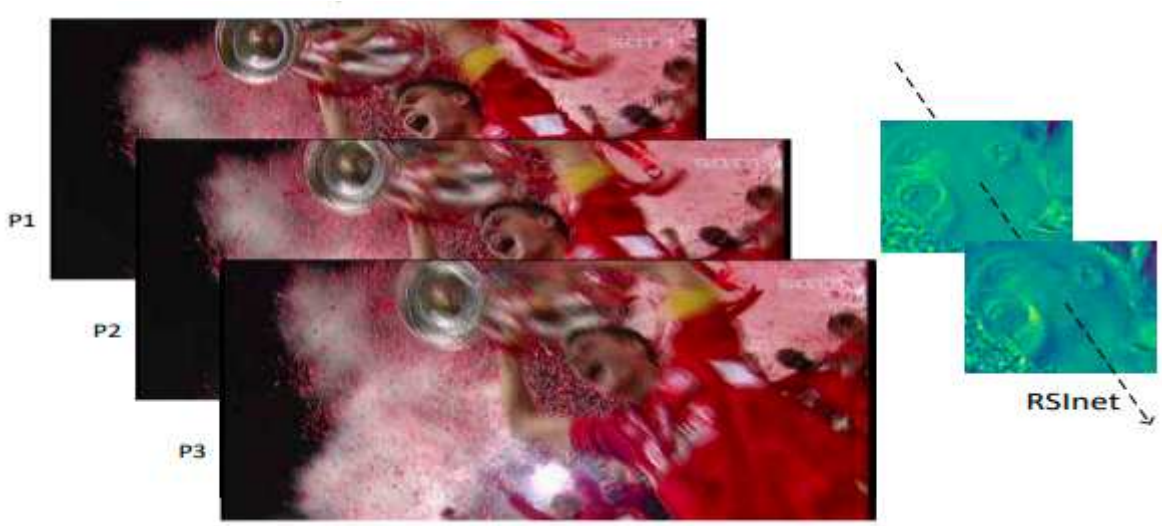

Figure 3

RELATIVE SPATIAL INFORMATION OF TARGETS IN CONTINUOUS SEQUENCE

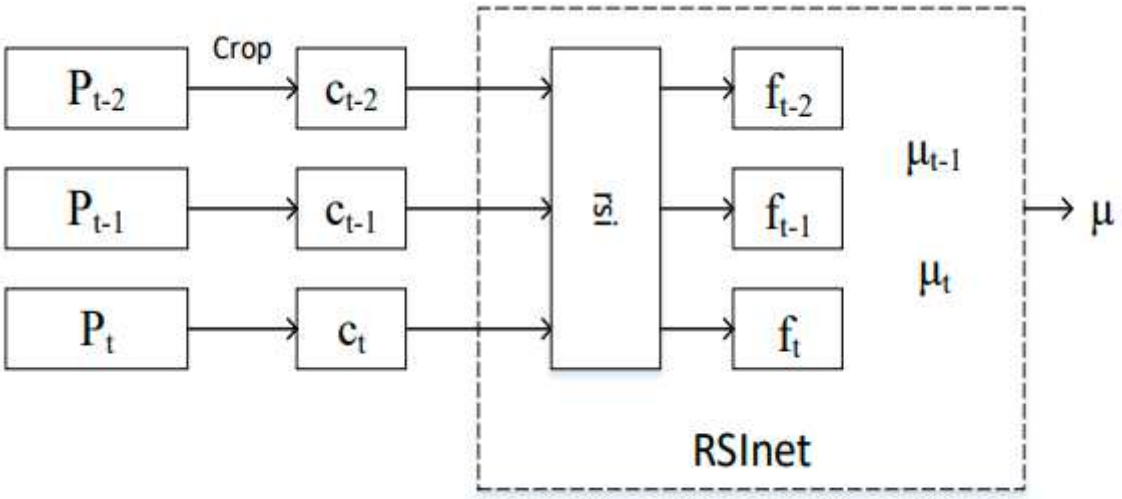

Figure 4

RSInet ALGORITHM STRUCTURE DIAGRAM

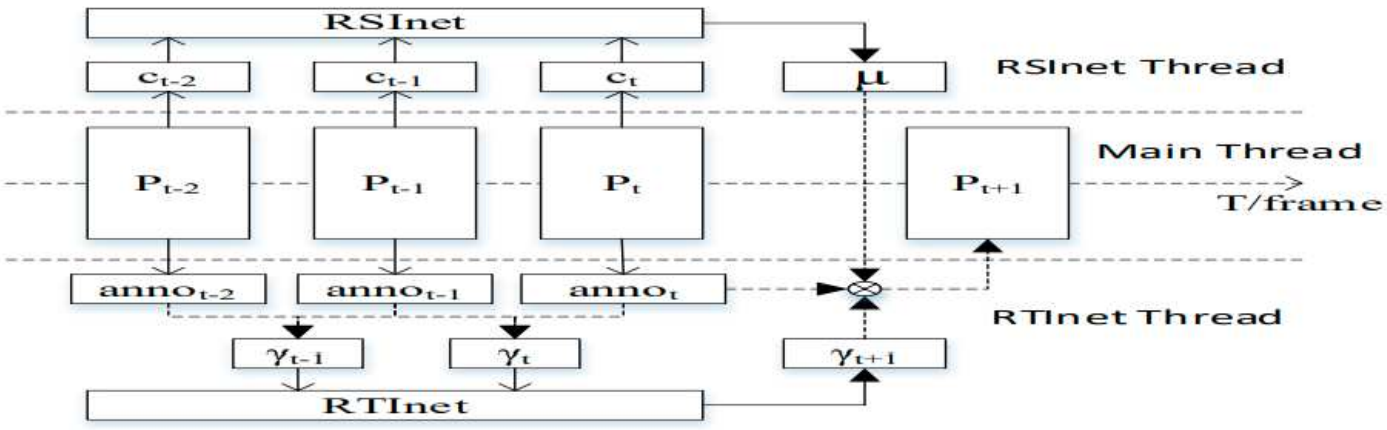

Figure 5

RTSnet ALGORITHM STRUCTURE DIAGRAM

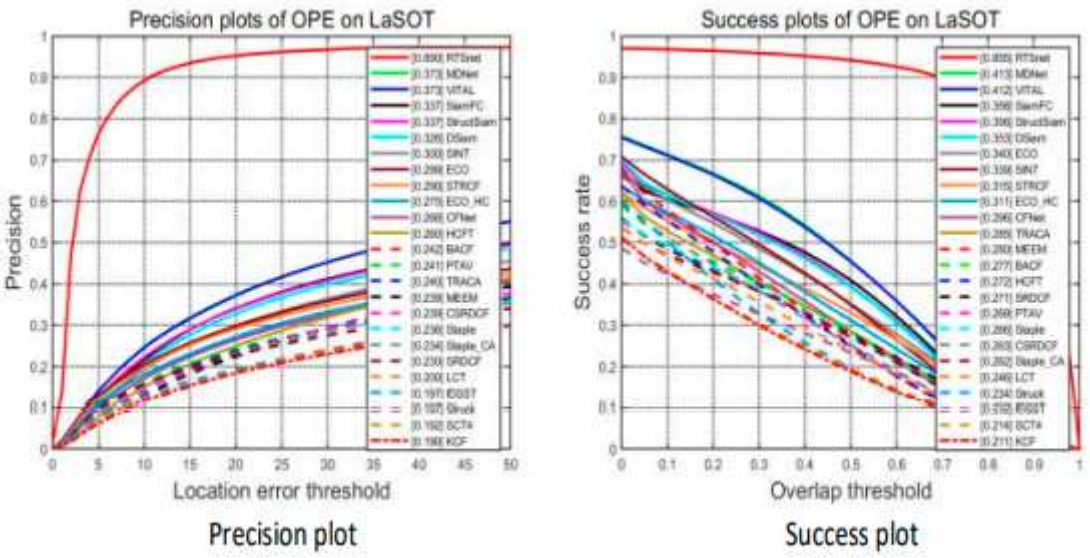

Figure 6

THE COMPARISON RESULTS OF RTSnet AND SOME STATE OF THE ART TRACKING NETWORKS UNDER THE LASOT TOOLKIT.

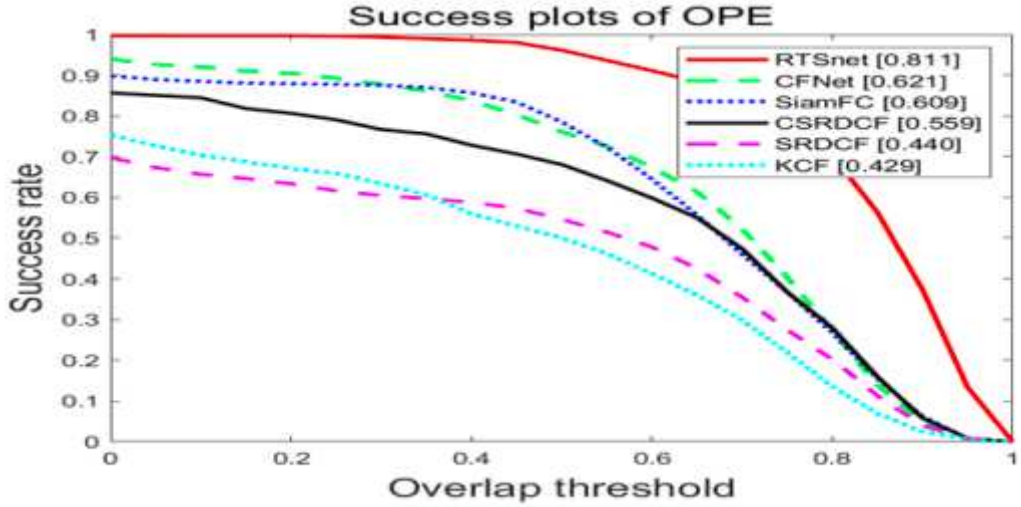

Figure 7

RTSnet'S SUCCESS RATE GRAPH UNDER THE OTB100 DATA SET

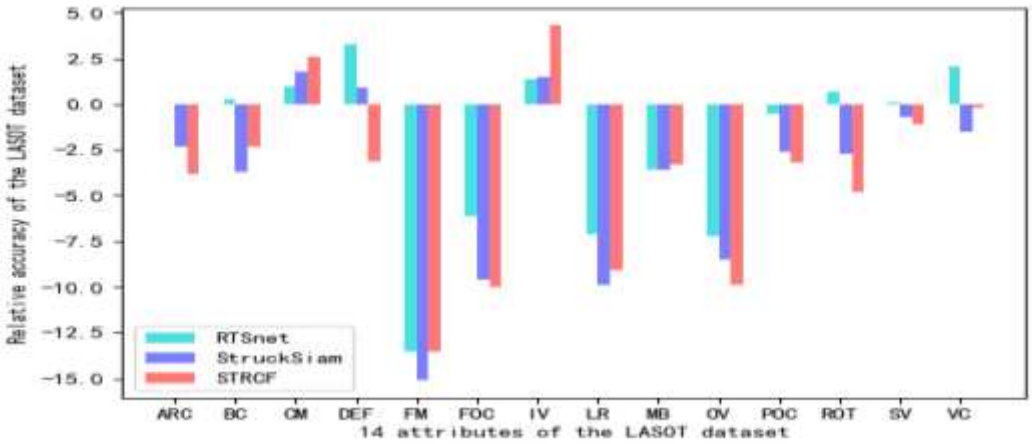

Figure 8

UNDER THE 14 ATTRIBUTES OF THE LASOT DATA SET, THE RELATIVE ACCURACY COMPARISON CHART OF RTSnet, STRUCKSIAM AND STRCF

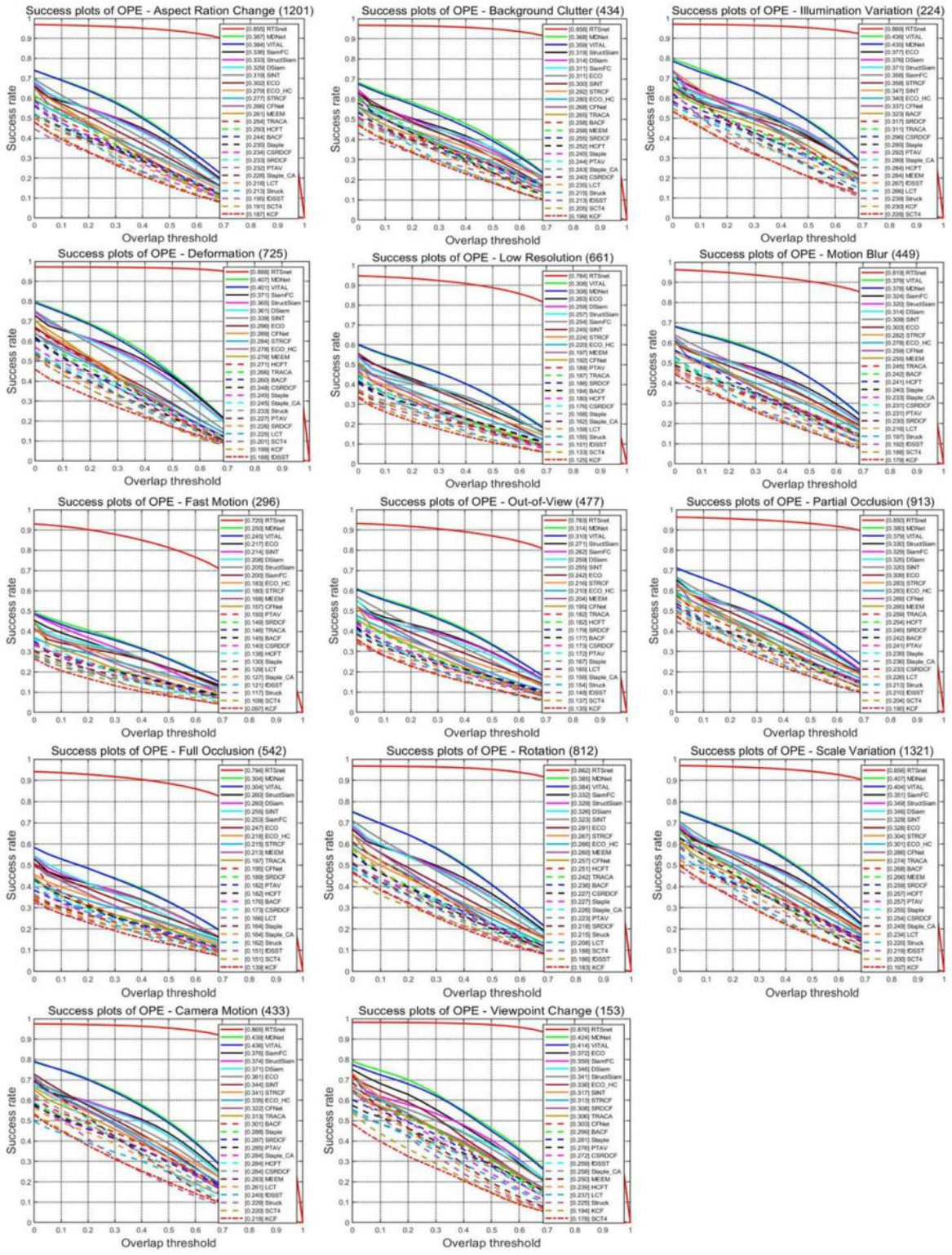

Figure 9

UNDER THE 14 ATTRIBUTES OF LASOT, THE ACCURACY COMPARISON CHART OF RTSNET AND SOME ALGORITHMS

Figure 10

IN 4 DIFFERENT SEQUENCES, THE TRACKING EFFECT COMPARISON CHART OF RTSnet, STRUCKSIAM AND STRCF