The Visual Computer

A Multivariate Intersection over Union of SiamRPN Network for Visual Tracking

--Manuscript Draft--

Manuscript Number:	
Full Title:	A Multivariate Intersection over Union of SiamRPN Network for Visual Tracking
Article Type:	Original article
Keywords:	Visual tracking; Multivariate Intersection over Union; Scale invariance; SiamRPN
Corresponding Author:	Jin Zhan, Ph.D Guangdong Polytechnic Normal University CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Guangdong Polytechnic Normal University
Corresponding Author's Secondary Institution:	
First Author:	Zhihui Huang
First Author Secondary Information:	
Order of Authors:	Zhihui Huang
	Huimin Zhao, Ph.D
	Jin Zhan, Ph.D
	Huakang Li
Order of Authors Secondary Information:	
Funding Information:	National Natural Science Foundation of China (62072122) — Dr. Huimin Zhao
	National Natural Science Foundation of China (61772144) — Dr. Jin Zhan
	Education Dept. of Guangdong Province (2019KSYS009) — Dr. Huimin Zhao
	Foreign Science and Technology Cooperation Plan Project of Guangzhou Science Technology and Innovation Commission (201807010059) — Dr. Jin Zhan
Abstract:	Although the SiamRPN algorithm perform well in visual tracking, it is easy to drift in occlusion and non-overlapping cases because the baseline L_1-smooth loss is sensitive to the scale of the bounding box. In this paper, we propose a Multivariate Intersection over Union (MIOU) regression target tracking algorithm based on SiamRPN network. We introduce a loss function with better generalization, which not only focuses on the overlap rate between the target box and predicted box, but also considers the distance of the central points and the shape in the bounding regression. Consequently, the loss metric of our method has scale invariance and which can accelerate the convergence speed and keep long-term tracking. By a series of quantitative and qualitative analysis, extensive experiments on benchmark datasets OTB2015 demonstrate that the proposed tracker is more robust to the challenges of occlusion, illumination change and fast motion.

The Visual Computer manuscript No.
(will be inserted by the editor)

A Multivariate Intersection over Union of SiamRPN Network for Visual Tracking

Zhihui Huang · Huimin Zhao · Jin Zhan✉ · Huakang Li

Received: date / Accepted: date

Abstract Although the SiamRPN algorithm perform well in visual tracking, it is easy to drift in occlusion and non-overlapping cases because the baseline ℓ_1-smooth loss is sensitive to the scale of the bounding box. In this paper, we propose a Multivariate Intersection over Union (MIOU) regression target tracking algorithm based on SiamRPN network. We introduce a loss function with better generalization, which not only focuses on the overlap rate between the target box and predicted box, but also considers the distance of the central points and the shape in the bounding regression. Consequently, the loss metric of our method has scale invariance and which can accelerate the convergence speed and keep long-term tracking. By a series of quantitative and qualitative analysis, extensive experiments on benchmark datasets OTB2015 demonstrate that the proposed tracker is more robust to the challenges of occlusion, illumination change and fast motion.

Keywords Visual tracking · Multivariate Intersection over Union · Scale invariance · SiamRPN

Z. Huang · H. Zhao · J. Zhan · H. Li
School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

Z. Huang
E-mail: zhihuihuanggd@foxmail.com

H. Zhao
E-mail: zhaohuimin@gpnu.edu.cn

J. Zhan✉
E-mail: gszhanjin@gpnu.edu.cn

H. Li
E-mail: lihuakang2020@163.com

1 Introduction

Visual target tracking is a subtask of computer vision and many advanced methods have been explored in this research area. It has numerous applications in many domains, including visual navigation, intelligent video surveillance system, intelligent human-computer interaction, medical diagnosis. Deep learning demonstrates powerfulness in extracting and processing semantic features, and can model the appearance of object by learning multimedia information. Inspired by this, many successful applications of deep learning have been achieved in computer vision, such as image segmentation, object detection, image classification, target tracking and so on.

Since 2013, the deep learning framework represented by SAE (Stack Autoencoding) [1,2], CNN (Convolution Neural Network)[3,4] and Siamese [5,6]has become the main backbone network of tracking algorithm. Deep Learning has been showing great success on object tracking. DLT [1] for the first time introduced deep network to break the bottleneck of traditional tracking model. After that CNN has been brought to enhanced the target learning capability of tracker[3], owing to the invariance principle in nonlinear changes such as translation, scale change and rotation. With the continuous research of depth structure, Tao et al.[5]successfully applied Siamese network as the backbone network of tracking algorithm, and made greatly progress in speed. However, it is weakness in many practical applications due to challenges such as illumination changes, partial occlusion, motion blur and low resolution, which obstruct the robust of tracking model. In recent years, the optimization trend of video target tracking focus on deepening neural network and enhancing feature extraction strategy, while ignoring the key role of loss function in model optimization. In computer vision such as target detection, recognition and semantic segmentation, the loss function can measure the performance of training model by comparing the difference between predicted model and actual data.

Inspired by the above observation, this paper introduces a Multivariate Intersection over Union (MIOU) as the new regression metric based on SiamRPN, where multivariate contains overlap rate, central point distance and shape. MIOU regression reflects both the actual position of the two bounding boxes and the aspect ratio by introducing the central point and shape information, which speed up the convergence rate of the bounding box and improve the generalization ability of the tracking model. In addition, our tracker is more robust in the case of the target bounding box surrounding the prediction box, and it is a relatively complete method of visual tracking regression measurement. Extensive experiments on benchmark datasets are carried out to validate our method effectiveness.

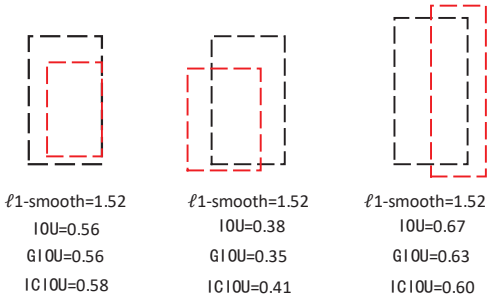
2 Related work

It is difficult to design a tracking model with both strong robustness and high precision. Therefore, many theoretical methods have been introduced to solve the tracking problem, such as classifier[7–9], sparse representation[10–12], saliency detection[13,14], feature selection[15–18] and deep learning[19, 20]. Prior depth trackers base on offline training and online fine-tuning achieve better results than traditional methods, and the online fine-tuning timely adjustment parameters to adapt the change of target better. However, despite the favorable performance of deep learning on object tracking, it is still limited by insufficient training samples, the foreground-background class imbalance, and high computational complexity in terms of time and space. Zhou et al.[19] used online AdaBoost to strengthen the classifier, and the updated strategy is robust to the appearance changes of object, which greatly improved the experimental accuracy. In the offline pre-training stage, MDNet [20]based on the multi-domain learning framework, through sharing layer and specific branch layer to enrich the feature of the target, and achieved excellent experimental results. Despite the great progress on algorithm accuracy, online depth methods are hardly meeting the requirements of real-time tracking.

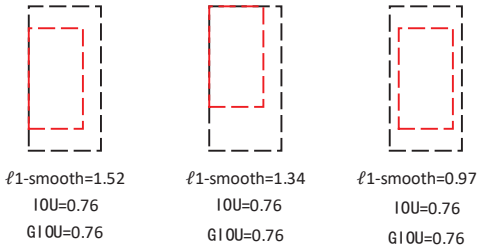
With the goal of achieving speed and accuracy at the same time, the Siamese network based on offline end-to-end training is introduced into the visual tracking. The algorithms based on Siamese network intend to learn a matching function from external data, and find the candidate patch that matches the target in the template frame by the learned matching function. Noting that it can achieve real-time tracking without model updating or online fine-tuning. Siamfc [21]uses the Siamese structure which is simple but effective as the core framework, and makes full convolution matching in the detection frame according to the template frame. The tracking speed reaches 86 fps, which has aroused widespread concern and accelerated the application of Siamese network in object tracking. In order to address the weakness of model robustness, Siamfc + +[22] proposed four guidelines: decomposition of classification and state estimation, non-ambiguous scoring, prior knowledge-free and estimation quality assessment, which effectively improved the generalization of the tracker. GCT (graph convolutional tracking) [23]constructs a graph convolution tracking framework base on the siamese structure, which acquires more sufficient and stable characteristic from detection frame by combining the temporal and spatial context information, and the experimental results show that the accuracy is improved greatly. SiamRPN [24]integrates target classification and regression positioning strategies, successively extracts features and generates candidate target regions using Siamese network and region proposal network, which effectively mitigate the issue of high computation complexity in multi-scale regression of Siamfc. However, SiamRPN is vulnerable to the case of object occlusion, background clutters and motion blur.

The tracking algorithm based on deep learning has been trying to deepen the network, but failed to address the problems of loss function. To alleviate the problem of class imbalance, Vital[25]adopts a high-order cost-sensitive loss

to decrease the effect of easily negative samples successfully. Therefor,designing a loss function with high generalization has certain theoretical significance and research value.



(a) The comparison of three loss metrics under different bounding scales.



(b) The comparison of three loss metrics under the case when real bounding box surrounds the predicted box.

Fig. 1: Due to ℓ_1-smooth is sensitive to the scale of bounding box, which cause failing to reflect the intersection information between real box(black) and predicted box(red) in the same value. Moreover, GIOU loss is transformed to IOU loss when real box (black) surrounds the predicted box (red), owing to the heavily relying on intersection over union.

In recent years, the research of target tracking algorithm is reviewed ℓ_n-normal form represented by ℓ_1-smooth function has become the mainstream loss metric, but it is sensitive to the scale change of the bounding box, which make it fails to optimize the case of non-overlapping bounding boxes, and easy to cause tracking failure. As shown in Figure 1 (a), the same ℓ_1-smooth loss value is infeasible to reflect the actual position and aspect ratios information between real box and predicted box. In order to further enhance the generalization performance of regression algorithm, Intersection over union (IOU) loss[26], Generalized intersection over union (GIOU) loss[27] are proposed suc-

cessively. According to the deficiency of IOU loss, GIOU loss can solve the issue of non-overlapping case, but the convergence speed is delayed due to strong dependence on intersection over union. When the real bounding box surrounds the predicted box, GIOU loss is equivalent to IOU loss, even keep the one value in different intersection case in Figure 1 (b). [28] proposed Distance intersection over union (DIOU) that addresses the problem of slow convergence by directly minimizing distance between central points of two bounding boxes. Furthermore, [28]also considers three important geometric variates in complete intersection over union (CIOU) loss. In designing shape parameter, CIOU uses square of angle difference between target box and predicted box to measures the consistency of aspect ratio, which ignores the vector quantity about angular quantity. In this paper, we summarize the vector of angular difference and the area of predicted bounding box to improve shape variate. Moreover, the improved shape variate is applied in MIOU loss, and we achieve notable performance gain by incorporating MIOU into SiamRPN.

The paper is organized by four sections. Firstly, we sorted out the research background in the "Introduction" section, and the related work is reviewed in second section. Afterwards, the "Proposed Method" section describes the our method in detail, including construction of network, design of new geometry metrics, target class and target locating. The "Experimental Results" section gives experiments process and results. Finally, the summary and future work were discussed in the "Conclusion and Future Work" section.

3 Proposed Method

Base on SiamRPN, the framework of our method contains Siamese subnetwork and Region Proposal subnetwork (RPN), where RPN network is constructed by two branches: classification and regression. Since the RetsNet50 pays more attention to the rich semantic information, we use RetsNet50 instead of Alexnet as the backbone network to break the spatial invariance restriction of Siamese subnetwork, which improve the capacity of target identification and helps the tracker adapt to the appearance varies better. In addition, we propose a multivariate intersection over union as the new regression metric in the RPN subnetwork to make the tracker more accurate. Figure 2 shows our method network structure diagram.

3.1 SiamRPN Framework

The SiamRPN is composed of Siamese subnetwork and RPN subnetwork. It abandons the traditional multi-scale test and online tracking, which makes the tracking speed very fast, with the maximum speed of 160 FPS. The RPN subnetwork uses multi-dimensional features to quickly generate target proposal regions and obtains K anchor points according to different preset aspect ratio.

The introduction of RPN keeps the network from suffering multi-scale regression calculation in target tracking, which not only improves the speed, but also makes the tracking results more accurate.

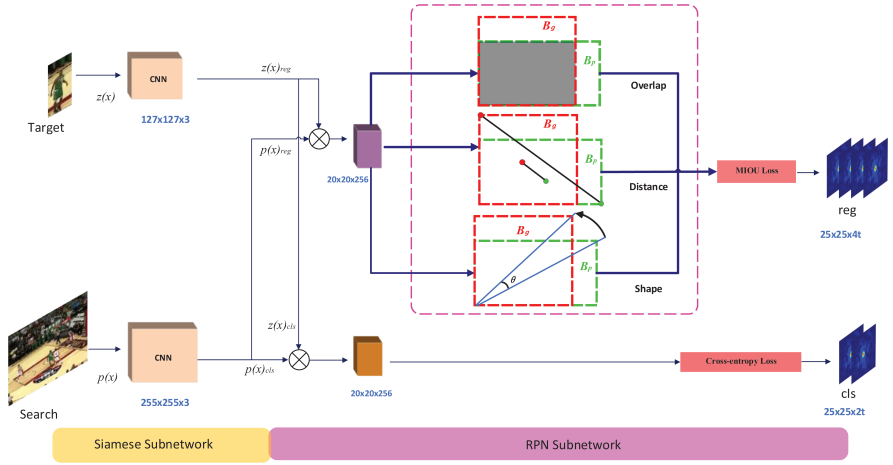


Fig. 2: Method network structure diagram.

As shown in Figure 2, the target frame $Z(x)$ and the search region $P(x)$ are input into the two subnetworks of Siamese module respectively. Meanwhile, they share the weights with the same structure during training. Considering the difference between sample classification and location function, RPN is further divided into classification branch (cls) and regression branch (reg). So the outputs of Siamese subnetwork are fed into cls and reg individually. In detail, the classification branch convolutes $p(x)_{cls}$ with $z(x)_{cls}$ as convolution kernel, and the output channel number of $A^{cls}_{w\times h\times 2t}$ is 2t, which indicates the positive and negative probability of candidate samples. Similarly, $z(x)_{reg}$ and $p(x)_{reg}$ produce the sensor $A^{reg}_{w\times h\times 4t}$ of 4t channels after correlation operation. We refer the regression result (d_x, d_y, d_w, d_h) as the four coordinates offsets of candidate targets. The specific operation process of the two tasks is as follows:

$$A^{cls}_{w\times h\times 4t} = z(x)_{cls} \otimes p(x)_{cls}, \tag{1}$$

$$A^{reg}_{w\times h\times 4t} = z(x)_{reg} \otimes p(x)_{reg}. \tag{2}$$

3.2 Multivariate Intersection over Union

Since loss function evaluates the difference between the predicted value and the real value, it plays an important role in generating observation model.

In addition, the smaller the loss function, the better the performance of the observation model. Therefore, a loss metric with good robustness is worth to be designed studiously. In SiamRPN, ℓ_1-smooth loss ignores the correlation of the four coordinate points of bounding box, so it is infeasible to reflect the scale information, as the Figure 1 shows that ℓ_1-smooth loss keeps the same value even the intersection between the two arbitrary boxes may be quite different. We observed that there are three critical factors in bounding box regression: the overlap rate, central point distance and shape. Aiming to address the above issue, a new metric is introduced in our method, which combines the three influence variates completely.

In Figure 3, Let B_g and B_p represent the target box and prediction box respectively, the position of the box is consisted of the coordinates of the two vertices in the lower left corner and the upper right corner, which $(\tilde{x_1}, \tilde{y_1}, \tilde{x_2}, \tilde{y_2})$ denotes the position of B_g , and (x_1, y_1, x_2, y_2) is represented by B_p . In additon, b_g and b_p are the center points of B_g and B_p ,and ρ represents the Euclidean distance of them. Noting that B_C denotes the smallest convex shape of B_g and B_p , c denotes the diagonal Euclidean distance of B_C.

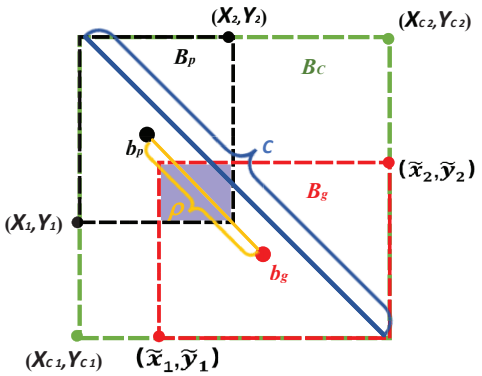


Fig. 3: Schematic diagram of target box and predicted box.

The coordination $(X_{C1}, Y_{C1}, X_{C2}, Y_{C2})$ of B_C is as follows:

$$X_{C1} = \min(\tilde{x_1}, x_1), Y_{C1} = \min(\tilde{y_1}, y_1), \tag{3}$$

$$X_{C2} = \max(\tilde{x_2}, x_2), Y_{C2} = \max(\tilde{y_2}, y_2). \tag{4}$$

We use I to denotes the maximum intersection box between B_g and B_p , the coordinate $(X_{I1}, Y_{I1}, X_{I2}, Y_{I2})$ of I come from the following formula:

$$X_{I1} = \max(\tilde{x_1}, x_1), Y_{I1} = \max(\tilde{y_1}, y_1), \tag{5}$$

$$X_{I2} = \min(\tilde{x_2}, x_2), Y_{I2} = \min(\tilde{y_2}, y_2). \tag{6}$$

A_{iou} means the iou value between two arbitrary boxes, the calculation process is as following:

$$A_g = (\tilde{x_2} - \tilde{x_1}) \times (\tilde{y_2} - \tilde{y_1}), \tag{7}$$

$$A_p = (x_2 - x_1) \times (y_2 - y_1), \tag{8}$$

$$A_I = (X_{I2} - X_{I1}) \times (Y_{I2} - Y_{I1}), \tag{9}$$

$$A_u = A_g + A_p - A_I, \tag{10}$$

$$A_{iou} = \frac{A_I}{A_u}, \tag{11}$$

where A_g and A_p are the area of B_g and B_p respectively, A_I means the area of intersection between B_g and B_p , and A_u is the area formed by the union of two bounding boxes. Hence, the loss function based on Multivariate Intersection over Union is as follows:

$$L_{miou} = 1 - A_{iou} + R(B_g, B_p) + S(B_g, B_p). \tag{12}$$

The third term in above equation is the penalty term of the central point distance, and the fourth one is shape adjustment term.

The mean of α is the balance factor, and δ is used to evaluate the aspect ratios alignment of the bounding box. Consequently, the scale information of candidate samples is adjusted according to the target aspect ratio. Then the symbols are defined as follows:

$$R(B_g, B_p) = \frac{\rho^2}{c^2}, \tag{13}$$

$$S(B_g, B_p) = \alpha\delta, \tag{14}$$

$$\delta = \frac{8}{\pi^2} \times \theta \times (w_p \times h_p), where \quad \theta = \theta_p - \theta_g, \tag{15}$$

$$\alpha = \frac{\upsilon}{1 - A_{iou} + \upsilon}, where \quad \upsilon = \frac{4\theta^2}{\pi^2}. \tag{16}$$

In Eq.(15), θ reflects the difference of aspect ratio between B_g and B_p, as shown in Figure 4. And θ_g denotes the inclination angle of the target box, while θ_p represents the predicted box inclination angle. Let $\theta_g = \arctan\frac{w_g}{h_g}$ and $\theta_p = \arctan\frac{w_p}{h_p}$, where w_g and h_g are taken from the scale of the target box, w_p and h_p represent the width and height of the predicted box. In order to achieve the aspect ratio alignment between B_g and B_p, we can see that $\theta < 0$ when $\theta_g > \theta_p$. Therefore, δ in Eq. (15) get the result less than zero, which means predicted box B_p rotates counterclockwise during regression optimization. On

the contrary, B_p rotates clockwise when $\theta > 0$, since $\theta_g < \theta_p$. This optimization process in bounding box shape can be visualized by Figure 4.

Algorithm 1: Multivariate Intersection over Union metric as bounding box loss

Input: ground truth B_p and predicted B_p bounding box

Coordinates: $B_g = (\tilde{x_1}, \tilde{y_1}, \tilde{x_2}, \tilde{y_2}), B_p = (x_1, y_1, x_2, y_2)$

Output: L_{miou}

1. Ensuring B_p meets the condition: $x_2 > x_1, \quad y_2 > y_1$:

$x_1 = \min(x_1, x_2), x_2 = max(x_1, x_2)$,

$y_1 = \min(y_1, y_2), y_2 = max(y_1, y_2)$,

2. Calculating area of B_g and B_p in Eq.(7) and Eq.(8), getting A_g, A_p.

3. Finding the coordinates of smallest enclosing box B_C in Eq.(3) and Eq.(4), getting

$(X_{C1}, Y_{C1}, X_{C2}, Y_{C2})$.

4. Calculating the diagonal Euclidean distance of B_C:

$c^2 = (X_{C2} - X_{C1})^2 + (Y_{C2} - Y_{C1})^2$

5. Calculating the center point of B_g and B_p, $b_g = (x_{b_g}, y_{b_g}), b_p = (x_{b_p}, y_{b_p})$:

$x_{b_g} = \frac{\tilde{x_1}+\tilde{x_2}}{2}, y_{b_g} = \frac{\tilde{y_1}+\tilde{y_2}}{2}$

$x_{b_p} = \frac{x_1+x_2}{2}, y_{b_p} = \frac{y_1+y_2}{2}$

6. Calculating the Euclidean distance between b_g and b_p:

$\rho^2 = (x_{b_p} - x_{b_g})^2 + (y_{b_p} - y_{b_g})^2$

7. Calculating the distance penalty term $R(B_g, B_p)$:

$R(B_g, B_p) = \frac{\rho^2}{c^2}$

8. Finding the coordinates of intersection I between B_g and B_p in Eq.(5) and Eq.(6),getting:

$(X_{I1}, Y_{I1}, X_{I2}, Y_{I2})$

9. Calculating area of I in Eq.(9), getting A_I.

10. Calculating A_{iou} in Eq.(10) and Eq.(11).

11. Calculating scale penalty term $S(B_g, B_p)$:

$S(B_g, B_p) = \alpha\delta$, where δ and α were calculated in Eq.(15) and Eq.(16).

12. $miuo = A_{iou} - R(B_g, B_p) - S(B_g, B_p)$

13. $L_{miou} = 1 - miou$

In contrast to ℓ_1-smooth baseline, our method not only focuses on the overlap variate, but also plays a distance and scale measure. Hence, it can

reflect the relative position relationship between the target and the candidate patch, and leverage the distance and shape information to guide the bounding box regression. We directly optimize the normalized distance between the two arbitrary boxes, which speeds up convergence during training phase. Moreover, we take advantage of aspect ratios alignment to avoid invalid regression in non-overlap or the real box contains the predicted box completely, it performs well in scale invariance by enhance robust of tracker. The pseudo code of the proposed algorithm is given in Algorithm1.

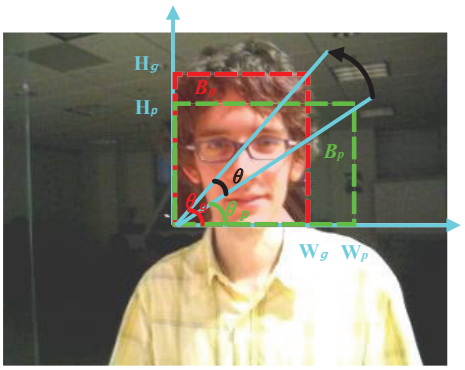


Fig. 4: Schematic diagram of optimization in bounding box shape.

4 Experimental Results

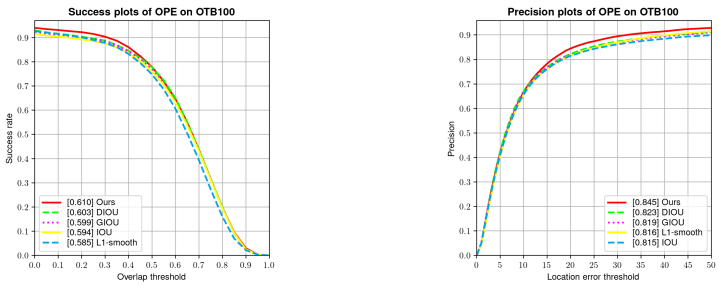
In this paper, we use ILSVRC-VID dataset to train model, which contains 30 basic categories. The dataset includes 3862 snippets for training, 555 snippets for verification, and 937 snippets for testing. It is also worthy to note that the training data includes different challenging scenes in the video sequence, such as illumination change, scale variation and low resolution, and provides the ground truth information of the tracking target. During training, for each video sequence, the target that comes from the first frame and the subsequent frame are input into template branch and search branch respectively. Among them, the template branch adjusts the input image block size to $127 * 127$ by using the convoluted operation, while the uniform scale of image block in searching branch is $255 * 255$. Finally, according to the overlapping rate calculation results, when iou value of candidate patches are greater than 0.6, it is judged as positive sample, while the iou value of negative samples are set to be no more than 0.3. The learning rate is initially set to and the number of anchors is 5. Since the target deformation difference is not obvious in the tracking process, the anchor ratios are set to (0.33,0.5,1,2,3), while the anchor area is constant, and a total of 20 epochs are performed. The experimental environment is dual GPU, 8g memory and NVIDIA GTX 1080gpu.

Table 1: The average center location error of ℓ_1-smooth, IOU, GIOU, DIOU and MIOU

Sequence	***Mhyang***	***Vase***	***Subway***	***Trellis***	***Jumping***	***Deer***	***Biker***	***Car4***	***David2***	***BlurFace***
ℓ_1-smooth	5.01	19.11	133.74	6.48	6.03	28.29	2.83	2.22	5.16	17.64
IOU	5.20	18.24	95.80	6.73	6.19	88.63	2.52	2.08	4.54	19.56
GIOU	7.43	14.23	3.82	8.22	5.93	29.06	67.88	2.45	4.46	16.89
DIOU	5.14	17.57	3.99	6.77	6.00	22.62	2.57	2.34	5.02	18.93
MIOU-(ours)	**3.99**	**13.71**	**3.07**	**5.90**	**5.66**	**14.11**	**2.27**	**1.87**	**4.07**	**13.98**

In the performance evaluation, we mainly compare our method against the four state-of-the-art metrics including ℓ_1-smooth loss, IOU loss, GIOU loss and DIOU loss simultaneously on OTB2015[29] benchmark dataset which consists of 100 fully annotated videos with various challenging attributes. We choose the average center location error as evaluation standard to quantify the performance of the methods. When the tracking effect of the tracker is better, the error value is lower, otherwise, the higher. To quickly validate the effectiveness of our proposed method, we only select 10 videos sequences from OTB2015 dataset. Table 1 shows the center error values in 10 videos, in which bold represents the best verification results. According to the results, our method performs better than ℓ_1-smooth,IOU,GIOU and DIOU.

With the goal of further certifying our algorithm, the performances of different approaches on the OTB2015 dataset are evaluated in precision plots and success plots. To increase the credibility of the results, we set the error threshold of 20 pixels in precision plots, and the area under curve values of success plots represents the overlap rate between the predicted box and target bounding box. As show in Figure 5, our method produces leading results in success and location precision, respectively.



(a) Success (b) Precision

Fig. 5: Quantitative comparisons of different metrics on OTB2015

Table 2: The effects of three variates on regression, including overlap rate, central point distance and shape.

Method	Success	Precision	Epoch
SiamRPN	0.585	0.816	15
SiamRPN-overlap	0.594	0.815	13
SiamRPN-overlap-distance	0.603	0.823	**12**
SiamRPN-overlap-shape	0.600	0.824	13
SiamRPN-overlap-distance-shape-(ours)	**0.610**	**0.845**	**12**

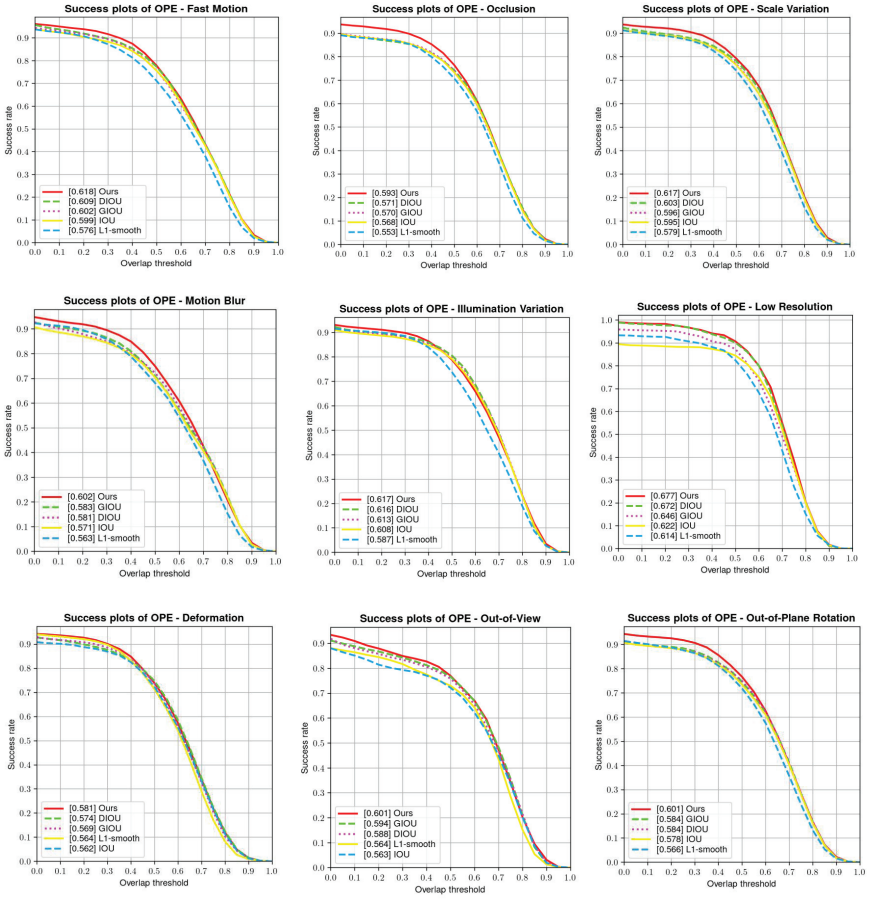
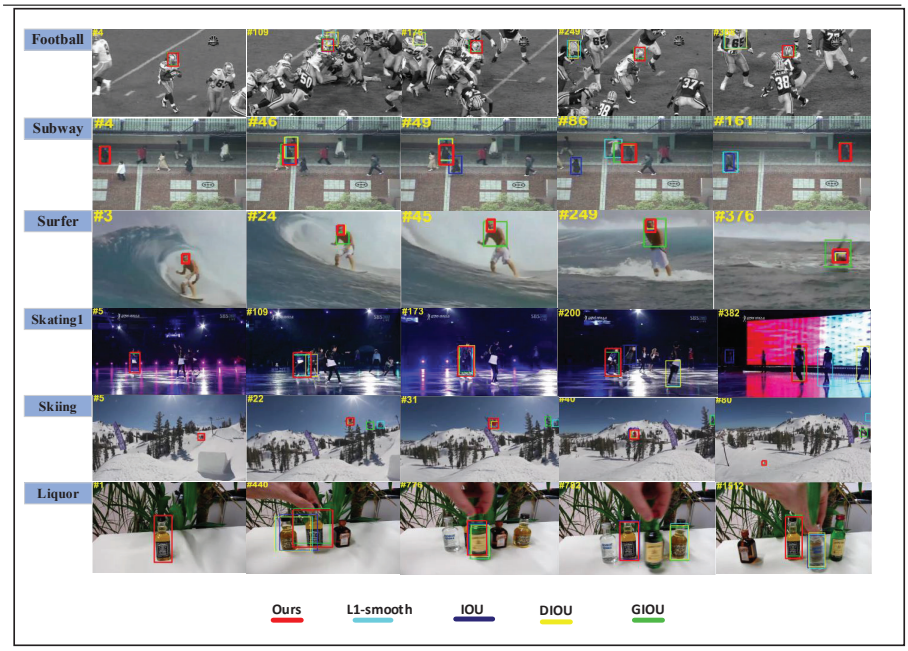


Fig. 6: The success plots between different algorithms in 9 sequence challenges, including Fast Motion, Occlusion, Scale Variation, Motion Blur, Illumination Variation, Low Resolution, Deformation, Out-of-View, Out-of-Plane.

In order to verify the effectiveness of three variates (overlap, central point distance and shape), we compare the precision score and success score of five algorithms by control variate. Table 2 illustrates the three geometric variates turn out to be helpful to improve SiamRPN tracking algorithm, but our method leads performance in success score and precision score. In "SiamRPN-overlap", we use overlap-based loss instead of baseline ℓ_1-smooth loss, which only considers the overlap rate variate in regression function. Meanwhile, "SiamRPN-overlap-distance" means applying overlap rate and the central point distance between B_p and B_g to design a regression function, while "SiamRPN-overlap-shape" denotes adopting overlap and shape geometric. As show in Table 2, it clearly indicates that our method which integrates three variates in bounding regression outperforms the others. It is also worthy to note that "Epoch" means the best iteration that the predicted box matches with the target box perfectly, the "Epoch" value smaller demonstrates the faster convergence. It is observed that the central point distance variate plays key work in speeding up convergence.

Figure 6 shows success plot of different algorithms on 9 video challenging attributes, including Fast Motion, Occlusion, Scale Variation, Motion Blur, Illumination Variation, Low Resolution, Deformation, Out-of-View and Out-of-Plane Rotation. Our method outperforms the other metrics trackers significantly in terms of all challenges, especially in occlusion, scale variation and motion blur, owing to provide regression direction in distance and shape of bounding box. Since GIOU loss and ℓ_1-smooth loss has strong laziness on intersection over union calculation, it shows slow convergence and easy divergence of training. However, our method is less sensitive to the 9 challenges, which performs more generalization and robustness.

We also illustrate the qualitative results in five different methods on a subset of 6 sequences in Figure 7. According to the OTB2015 benchmark, the sequences of Football and Subway contain serious OCC(Occlusion), DEF(deformation) and BC(Background Clutters), ℓ_1-smooth and DIOU occur tracking drift in ♯109 of Football ,but our method,IOU and GIOU keep tracking successfully in the end. Sequences Surfer is typical of target SV (Scale Variation) and FM (fast motion), as we can see that GIOU has the problem of serious scale tracking failure, while other trackers perform well in these challenges. In sequences Skating1, many trackers suffer from short-term occlusion in ♯173, however, our method and GIOU can effectively deal with the non-overlap case and reposition target in ♯200. Moreover, our approach can identify the target in noticeable illumination change. When IV (illumination change) and SV (Scale Variation) occur in Skiing and Liquor simultaneously, ℓ_1-smooth, IOU, DIOU and GIOU are seriously affected by the susceptibility of scale, and leading to tracking failed. But our method has a good tracking effect in these cases and maintain long-term tracking. In general, the results clearly show that using our method as the bounding box regression loss performs consistently better in videos, while some failure cases are occurred in ℓ_1-smooth loss, IOU loss, GIOU loss and DIOU loss.



(a) Success

Fig. 7: Qualitive results of the proposed method(red), ℓ_1-smooth loss(blue), IOU loss(cyan), GIOU loss(green) and DIOU loss(yellow) (Football, Subway, Surfer, Skating1, Skiing, Liquor) on OTB2015.

5 Conclusion and Future Work

SiamRPN algorithm has outstanding performance in the tracking algorithm based on Siamese network, but it is prone to drift in occlusion and non-overlapping case, and the baseline ℓ_1-smooth loss is sensitive to the scale of the bounding box. Therefore, this paper proposes a multivariate intersection over union regression target tracking algorithm based on SiamRPN. We design a loss function with better generalization, which not only focus on the overlap of the bounding box, but also considers the distance between the central point and the shape of the bounding box. Consequently, the new metric has scale invariance, provides distance and scale direction for the bounding box regression, respectively. In the meantime, our method speeds up the convergence in training. Moreover, it can maintain long-term tracking by solving the problem that ℓ_1-smooth loss tracking failure in non-overlapping case. Collecting experiment results on OTB2015 dataset, a series of quantitative and qualitative analysis show that the proposed method is more robust to the challenges of occlusion, scale change, illumination change and fast motion.

We will further investigate this work, firstly, it is sensitive to similar distractor of intra-class. We plan to integrate structure information of object to

enhance capability of discriminating distractors. Secondly, we figure out anchor shape has a great influence on the effectiveness of the model, so we will introduce adaptive feature fusion to refine features based on the underlying anchor shapes.

Acknowledgements This research was funded by National Natural Science Foundation of China (No.62072122, No. 61772144), and Education Dept. of Guangdong Province (No.2019KSYS009), Foreign Science and Technology Cooperation Plan Project of Guangzhou Science Technology and Innovation Commission (No.201807010059).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Wang, N.,Yeung, D. Y. Learning A Deep Compact Image Representation for Visual Tracking.In: Proceedings of the Neural Information Processing Systems (NIPS), pp.809-817(2013).
2. Zhou, X., Xie, L., Zhang, P,Zhang Y. An Ensemble of Deep Neural Networks for Object Tracking.In: Proceedings of 2014 IEEE International Conference on Image Processing (ICIP),pp.843-847(2014).
3. Wang, N., Li, S., Gupta, A., Yeung, D.Y. Transferring Rich Feature Hierarchies for Robust Visual Tracking. arXiv2015(2015).
4. Nam, H.,Han, B. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. arXiv 2016(2016).
5. Tao, R., Gavves, E., Smeulders, A.W.M. Siamese Instance Search for Tracking. In:Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.850-865(2016).
6. Xuan, S.,Li, S.,Zhao, Z., Kou, L.,Zhou, Z., Xia, G. Siamese Networks with Distractor-reduction Method for Long-term Visual Object Tracking, Pattern Recognit. 8 October 2020, 107698.
7. Grabner, H.,Leistner, C., Bischof, H. Semi-Supervised Online Boosting for Robust Tracking. In: Proceedings of European Conference on Computer Vision (ECCV),pp.234-247(2008).
8. Babenko, B., Yang, M.H., Belongie, S. Visual Tracking with Online Multiple Instance Learning. In:Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 983-990(2009).
9. Kalal, Z., Mikolajczyk, K., Matas, J. Tracking-Learning-Detection. IEEE Trans. Pattern Anal. Mach. Intell. 34(7). 1409-1422(2012).
10. Mei, X., Ling, H. Robust Visual Tracking Using l1 Minimization. In:Proceedings of IEEE International Conference on Computer Vision (ICCV). pp.1436-1443(2009).

11. Wang, D., Lu H., Yang, M.H. Online Object Tracking with Sparse Prototypes. IEEE Trans. Image Process (TIP).22(1). 314-325(2013).
12. Zhang, T., Liu, S., Xu, C., Yan S., Ghanem Be., Ahuja N., Yang M.H. Structural Sparse Tracking. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.150-158(2015).
13. Wang, Z., Ren, J., Zhang, D.,Sun M.,Jiang J. A Deep-Learning based Feature Hybrid Framework for Spatiotemporal Saliency Detection Inside Videos. Neurocomputing.287. 68-83(2018).
14. Yan Y.,Ren J.,Zhao H., Sun G., Wang Z., Zheng J., Marshall S.,Soraghan J. Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cognit. Comput. 10(1).94-104(2017).
15. Han, J., Zhang, D., Cheng, G., Lei G., Ren J. Object Detection in Optical Remote Sensing Images based on Weakly Supervised Learning and High-Level Feature Learning. IEEE Trans. Geosci. Remote. Sens. 53(6).3325-3337(2015).
16. Zabalza, J.,Ren, J., Zheng, J., Zhao H.,Qing C., Yang Z., Du, P., Marshall S. Novel Segmented Stacked Autoencoder for Effective Dimensionality Reduction and Feature Extraction in Hyperspectral Imaging. Neurocomputing.185.1-10(2016).
17. Tschannerl, J., Ren, J., Yuen, P., Sun, G., Zhao, H., Yang, Z., Wang, Z., Marshall S. MIMR-DGSA: Unsupervised Hyperspectral Band Selection based on Information Theory and A Modified Discrete Gravitational Search Algorithm. Inf. Fusion.51.189-200(2019).
18. Xia, H., Zhang Y.,Yang M., Zhao Y.Visual tracking via deep feature fusion and correlation filters. Sensors. 20(12).3370(2020).
19. Zhou, X., Xie, L., Zhang, P., et al. An Ensemble of Deep Neural Networks for Object Tracking. In:Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 843-847(2014).
20. Nam, H., Han, B. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. In:Proceedings of Computer vision and pattern recognition(CVPR). pp.4293-4302(2016).
21. Bertinetto, L., Valmadre J., Henriques, J. F., Vedaldi, A., Torr, Philip H.S. Fully-Convolutional Siamese Networks for Object Tracking. Proceedings of European Conference on Computer Vision (ECCV).pp.850-865(2016).
22. Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In:Proceedings of AAAI.pp.12549-12556(2020).
23. Gao, J., Zhang, T., Xu, C. Graph Convolutional Tracking. In:Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR).pp.4649-4659(2019).
24. Li, B., Yan, J., Wu, W., Zhu Z., Hu X. High Performance Visual Tracking with Siamese Region Proposal Network.In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp.8971-8980(2018).
25. Song, Y., Ma, C., Wu, X., Gong L., Bao L.,Zuo W., Shen C., Lau R.W.H., Yang M.H. VITAL: Visual Tracking Via Adversarial Learning. Proceedings of Conference on Computer Vision and Pattern Recognition. pp.8990-8999(2018).
26. Yu, J., Jiang, Y., Wang, Z., et al. UnitBox: An Advanced Object Detection Network.In: Proceedings of the 24th ACM international conference on Multimedia. pp.516-520(2016).

27. Rezatofighi, H.,Tsoi, N., Gwak, J. Y., Sadeghian A., Reid I.,Savarese S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.pp. 658-666. June 2019.
28. Zheng, Z., Wang, P., Liu, W., Li J., Ye R., Ren D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Proceedings of AAAI,pp.12993-13000(2020).
29. Wu, Y., Lim, J., Yang, M.H. Object Tracking Benchmark. IEEE Trans. Pattern Anal. Mach. Intell.37.1834–1848(2015).

Your PDF file "Manuscript.pdf" cannot be opened and processed. Please see the common list of problems, and suggested resolutions below.

Reason:

Other Common Problems When Creating a PDF from a PDF file

You will need to convert your PDF file to another format or fix the current PDF file, then re-submit it.

The Visual Computer manuscript No.
(will be inserted by the editor)

A Multivariate Intersection over Union of SiamRPN Network for Visual Tracking

Zhihui Huang · Huimin Zhao · Jin Zhan✉ · Huakang Li

Received: date / Accepted: date

Z. Huang · H. Zhao · J. Zhan · H. Li
School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

Z. Huang
E-mail: zhihuihuanggd@foxmail.com

H. Zhao
E-mail: zhaohuimin@gpnu.edu.cn

J. Zhan✉
E-mail: gszhanjin@gpnu.edu.cn

H. Li
E-mail: lihuakang2020@163.com

1 Author Biographies

Fig. 1: **Zhihui Huang**,received her B.S. degree in Computer Science and Technology from Guangdong Ocean University, Zhanjiang, in 2018. Currently, she is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. Her research interests include machine learning and computer vision.

Fig. 2: **Huimin Zhao**,was born in Shanxi, China, in 1966.He received the B.Sc. and the M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Sun Yat-sen University, Guangzhou, China, 2001. He is currently a Professor and the Dean with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou. His research interests include image, video, and information security technology.

Fig. 3: **Jin Zhan**,received the Master's and Doctor's degrees from Sun Yat-sen University,Guangzhou, China, in 2004 and 2015, respectively. She is currently an associate professor in the School of Computer Science, Guangdong Polytechnic Normal University,Guangzhou.Her research interests include image and video intelligent analysis, machine learning and computer vision.

Fig. 4: **Huakang Li**,received his B.S. degree in Electrical Engineering and Automation from Southern College of Sun Yat-sen University, Guangzhou, in 2019. Currently, he is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. His research interests include deep learning and gait recognition.

The Visual Computer manuscript No.
(will be inserted by the editor)

A Multivariate Intersection over Union of SiamRPN Network for Visual Tracking

Zhihui Huang · Huimin Zhao · Jin Zhan✉ · Huakang Li

Received: date / Accepted: date

Z. Huang · H. Zhao · J. Zhan · H. Li
School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

Z. Huang
E-mail: zhihuihuanggd@foxmail.com

H. Zhao
E-mail: zhaohuimin@gpnu.edu.cn

J. Zhan✉
E-mail: gszhanjin@gpnu.edu.cn

H. Li
E-mail: lihuakang2020@163.com

1 Author Biographies

Fig. 1: **Zhihui Huang**,received her B.S. degree in Computer Science and Technology from Guangdong Ocean University, Zhanjiang, in 2018. Currently, she is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. Her research interests include machine learning and computer vision.

Fig. 2: **Huimin Zhao**,was born in Shanxi, China, in 1966.He received the B.Sc. and the M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Sun Yat-sen University, Guangzhou, China, 2001. He is currently a Professor and the Dean with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou. His research interests include image, video, and information security technology.

Fig. 3: **Jin Zhan**,received the Master's and Doctor's degrees from Sun Yat-sen University,Guangzhou, China, in 2004 and 2015, respectively. She is currently an associate professor in the School of Computer Science, Guangdong Polytechnic Normal University,Guangzhou.Her research interests include image and video intelligent analysis, machine learning and computer vision.

Fig. 4: **Huakang Li**,received his B.S. degree in Electrical Engineering and Automation from Southern College of Sun Yat-sen University, Guangzhou, in 2019. Currently, he is a Master student in the School of Computer Science, Guangdong Technical Normal University, Guangzhou. His research interests include deep learning and gait recognition.