Research on Target Tracking Algorithm in occlusion scene

Jianguang Zhao¹*, Xinying Cao¹

¹HEBEI UNIVERSITY OF ARCHITECTURE, Zhangjiakou, Hebei, 075000, China
*Corresponding author’s e-mail: 12112074@bjtu.edu.cn

Abstract. Target tracking is widely used in video image, intelligent transportation, behavior analysis and other fields. At the same time, target tracking technology faces more and more challenges. In this paper, the target tracking technology in occlusion scene is studied, and the target tracking method based on spatiotemporal context and the target tracking method combined with depth network model are introduced respectively. The main ideas and principles of the two methods are analyzed, and the improvements based on these two methods are summarized. Finally, the problems existing in the future development of occluded target tracking algorithm are prospected.

1. Introduction

The gradual popularization of video surveillance meets people's needs for media, entertainment, security and other aspects. At present, video monitoring is developing towards intelligence. The non-human monitoring video can realize automatic accident detection, alarm and other functions, which is convenient for people's life and also provides protection for people's personal safety. Target tracking technology is the key technology of intelligent video surveillance, and also a hot research field of computer vision. However, in the process of target tracking, due to the influence of illumination environment, the phenomenon of target occlusion will appear, which seriously affects the detection and tracking of targets. Object occlusion detection and processing has become the most basic and challenging problem in the field of target detection.

Object occlusion can be divided into static occlusion and dynamic occlusion. Static object occlusion generally refers to the situation that the object is occluded or the moving object and background are relatively static in the target detection of static image. Static object occlusion processing is relatively simple, and the difficulty lies in dynamic object occlusion. Dynamic object occlusion means that the moving background and target are in non-stationary state. According to the size of occlusion area, the occlusion can be divided into local occlusion and severe occlusion. There are usually two ways to solve the problem of target occlusion: (1) when the target is partially occluded, the local matching can be carried out by using the unobstructed regional features to realize the target positioning and recognition. When the target is seriously occluded, each frame image of the video sequence is detected and recognized to realize the target tracking [1]. (2) The target information in the next frame can be predicted by using the unobstructed target in the previous frame to realize target tracking [2].

2. Occlusion target tracking based on Spatio-Temporal context

2.1. Spatio-Temporal context

In object tracking, there are a lot of temporal and spatial information between moving object and background, which is called local context. The tracking algorithm uses the local context to establish the
relationship between adjacent frames to achieve real-time and accurate target tracking. The spatial information contains a specific relationship between the target and the background, so as to distinguish the foreground from the background. When the moving target is occluded, the position of the target center relative to the background will not change dramatically. The location of the target can be located by using this feature. In the adjacent two frames of video, the moving object will not produce mutation, so the position of the moving object in the next frame can be predicted by the position of the moving object in the previous frame to realize the target tracking. The local context relation can be used to realize fast target tracking and has good robustness.

2.2. The step of algorithm implementation

Target tracking algorithm based on spatiotemporal context (STC) [3] is mainly divided into two steps: establishing a target detection model based on time and space; using the target information of the previous frame to predict the position of the target in the next frame.

2.2.1. Establish the detection model of time and space

The tracking algorithm mainly uses Bayesian total probability formula, uses conditional probability and prior probability to establish the local context relationship of the target, and finds the point with the maximum confidence through the confidence graph, which is the target location.

In essence, the tracking problem can be regarded as a confidence map (likelihood estimation) to calculate the estimated position $X$ of the target. Assuming that the target has been determined, in the current frame $t$ frame, the confidence map of the target is expressed as follows:

$$c_i(x) = P(x | o)$$

Where $x \in \mathbb{R}^2$ represents the position of the target, $O$ represents the target, and the position of the maximum point in the confidence graph is the target position.

The above formula is decomposed by using the formula of full probability:

$$c_i(x) = P_{c(z) \in X^*} P(x | c(z), o) P(c(z) | o)$$

Where $X^* = \{c(z) = (I(z), z) | z \in \Omega c(x^*)\}$ is the feature set of the context, $I(z)$ represents the pixel value of the target point, $X^*$ represents the center of the target, $\Omega c(x^*)$ represents the neighborhood of the target center. As can be seen from the above figure, the likelihood function can be divided into two parts: conditional probability function and prior probability function $P(x | c(z), o)$ represents the spatial information between the target and the background, $P(c(z) | o)$ represents the occurrence probability of local context.

The conditional probability in formula (2) can be expressed as follows:

$$P(x | c(z), o) = h^{se}(x - z)$$

$h^{se}$ is the relationship between the target and the spatial context, indicating the relative distance and direction between the target and the context. This function describes the relative spatial relationship of each feature point in the target. When the distance between the feature point and the target center point is the same, the position of the feature point relative to the center point can be distinguished by the direction, so as to determine the position of the whole target and reduce the occurrence rate of target false detection.

The prior probability in formula (2) can be expressed as follows:

$$P(c(z) | o) = I(z)w_a(z - x^*)$$

Where $w_a(z) = \frac{|z^2|}{\sigma^2}$ is a weighting function, the closer the point to $X$, the greater the weight. $A$ is the normalization constant and $\sigma$ is the scale parameter.

2.2.2. Predicting the position of the next frame

According to the previous step, the position of the target in the $T$ frame is obtained. Now the confidence map of the $T + 1$ frame is obtained by the known position of the $T$ frame:
\[ c_{t+1}(x) = P(x \mid o) = b e^{-\frac{|x-x^*|^\beta}{\alpha}} \]  

(5)

Where \( b \) is the normalization constant, \( \beta \) is the shape parameter and \( \alpha \) is the scale parameter.

Combined with formula (1) (2) (3) (4), we can get the following results:

\[
c_{t+1}(x) = \sum_{z \in \Omega} h^\infty (x - z) l(z) \omega_\sigma (z - x^*)
= h^\infty (x) \otimes (I(x) \omega_\sigma (x - x^*))
\]  

(6)

\( \otimes \) is the convolution operation, and finding the point with the highest confidence is the location of the target.

![STC algorithm implementation flow chart](image)

The context based object tracking algorithm has good robustness and real-time performance, and it can still track the target well in the situation that the target is occluded.

3. Occlusion target tracking combined with depth network model

3.1. Deep network model

Deep learning has become the most popular research field of artificial intelligence. Deep learning method has been widely used in target detection and target tracking. Convolution neural network is the basis of deep network model. Convolutional neural network is similar to human brain, which simulates...
the learning process of people by constructing convolutional neural network. The convolution neural network can be divided into input layer, convolution layer, pooling layer, full connection layer and output layer.

Figure.2 LeNet-5 network structure

LeNet-5 (as shown in Figure.2) is the earliest deep neural network, which is the first work of convolutional neural network. However, the effect of convolutional neural network was not good compared with the traditional classification algorithms such as SVM and PCA. After that, the real rise of convolutional neural network is AlexNet, which won the champion in ImageNet competition in 2012. Hinton added dropout layer to the network, which greatly improved the recognition effect of the network. After the continuous improvement of researchers, a variety of new convolutional neural networks have emerged, such as ResNet, siamese network, capsule network, full convolution neural network, and many detection frameworks such as Fast R-CNN, Faster R-CNN, Mask R-CNN, YOLO, etc.

3.2. Anti occlusion tracking principle combined with depth network model

The function of depth network model in tracking is feature extraction, so occlusion compensation strategy is introduced in target tracking by using the advantages of network model. According to the occlusion area, occluded objects can be divided into several categories (the specific ratio can be set by ourselves): no occlusion (0%), partial occlusion (0% - 50%), severe occlusion (50% - 100%) and complete occlusion (100%). The occluded area is marked according to the occlusion strategy by manually labeling the data set, and the target data set with occlusion mark is established. Then the appropriate detection framework is selected to train the labeled data, and different optimization strategies are used to reduce the feature loss of the target.

Figure.3 Schematic diagram of deep network tracking

4. Algorithm improvement research

4.1. Improved target tracking algorithm based on spatiotemporal context

Although the context object tracking algorithm can improve the detection rate of occluded objects, it also has some disadvantages in some aspects. Therefore, many researchers have improved the traditional context target tracking technology to improve the detection efficiency of the algorithm. When the video speed is fast, the context information of the target will change dramatically and it is difficult to locate.

Chu et al. [4] introduced the occlusion discrimination module to judge the occlusion situation. The occlusion discrimination module uses the bidirectional trajectory error method, uses the median flow tracker to infer the forward trajectory, and then infers the backward trajectory, the error between the two tracks becomes the bidirectional trajectory error, and uses the error value to judge whether the target is
occluded. For the seriously occluded target, we do not track directly, but choose to give up tracking and re-detect the target. This method can solve the problems of static occlusion and dynamic occlusion, improve the stability and detection accuracy of the algorithm, and reduce the complexity of the algorithm.

Liang et al. [5] used the combination of directional gradient histogram and grayscale image to replace the single gray feature in the traditional context target tracking confidence map, which improved the anti-interference performance of the algorithm, and improved the fixed learning rate problem in the traditional algorithm. In the tracking process, adaptive learning rate was used to update the context detection template according to the change of speed, which improved the tracking performance the accuracy of the trace.

### 4.2. Improvement of target tracking algorithm combined with deep network model

Chen et al. [6] Based on the siamese network, adopted the latest target tracking framework Siam-VGG network to realize target tracking under occlusion. Siamese network is a network composed of two identical or similar networks, and the two networks learn separately. Siamese network is often used in the scene of target matching and video tracking. In this paper, occlusion is divided into four categories: no occlusion, partial occlusion, total occlusion and target loss. Different anti occlusion strategies are defined by judging the size of network confidence map and connected domain, which provides a new research idea for target tracking in occlusion situation.

In the simple occlusion target classification mode, the new strategy was added by tie et al. [7], which divided occlusion markers into more complex situations: mutual occlusion, background occlusion and composite occlusion. Under each large occlusion mark, more detailed labeling was carried out according to the occlusion area. Different occlusion compensation coefficients are selected to reduce the target information loss under different conditions. The loss function is optimized by using the YOLO network model, which improves the accuracy of traditional YOLO target detection.

| METHOD | SPEED | ROBUSTNESS | REAL-TIME | ACCURACY |
|--------|-------|------------|-----------|----------|
| 4 refs | higher | higher | higher | higher |
| 5 refs | lower | higher | higher | higher |
| 6 refs | lower | higher | lower | higher |
| 7 refs | lower | higher | lower | higher |

### 5. Summary and Prospect

Generally speaking, the context based target tracking algorithm has advantages over the deep network target tracking algorithm in terms of speed, but there are still some deficiencies in accuracy. The training of neural network needs a lot of time, and also needs manual data annotation, there are many improvements. Although there are many methods for anti occlusion target detection, in the face of dense occlusion, the algorithm still can not achieve complete tracking, and tracking in rain, snow, fog weather is a more challenging problem.

### Acknowledgments

This work was financially supported by Zhangjiakou City's 2018 municipal-level science and technology plan financially-funded project. Project name: Research on Cloud Computing Data Center Network Architecture Optimization of Olympic Stadiums. Project Number: 1811009B-19, and the “Three, Three, Three Talents Project” of Hebei Province in 2018, the project name: Software-Defined Wireless Sensor Network Data Forwarding Research Project award number: A201803005. Hebei Technological Innovation Guidance Plan Project: Project Title: Research on High-precision Location Technology of Snow and Ice Accidents in VR Scene Based on 5G.Project Number: 20470302D.

### References

[1] ZHANG T Z, GHANEM B, LIU S, et al. Robust Visual Tracking via Multi-task Sparse
[1] Learning // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2012: 2042-2049.

[2] YANG H, ALAHARI K, SCHMID C. Online Object Tracking with Proposal Selection // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2015: 3092-3100.

[3] Zhang K H, Zhang L, Liu Q S, et al. Fast visual tracking via dense spatio—temporal context learning[C]ff Proc of Europe—an Conference on Computer Vlsion(ECCV 2014), 2014: 127—141.

[4] Chu Jun, Zhu Tao, Miao Jun, Jiang Landa. Object tracking algorithm based on occlusion detection and spatiotemporal context information. Pattern recognition and artificial intelligence, 2017,30 (8): 718-727.

[5] Liang Xiao, Shen Maoxian, Chen Guodong. An anti occlusion spatiotemporal context target tracking [J]. Firepower and command control, 2019, 44 (6): 54-58.

[6] Chen Fujian, Xie Weixin. Target tracking algorithm in siamvgg network with anti occlusion mechanism [J]. Signal processing, 2020,36 (04): 562-571.

[7] Tian Jun, Song Wei, Yin Fan, et al. Object detection algorithm based on occlusion markers [J]. Journal of Central South University for Nationalities (NATURAL SCIENCE EDITION), 2020,39 (3): 302-308.