Research on Multi-Target Stable Tracking Algorithm Based on Detection and Tracking Fusion

Shikai Wang\(^1\), Zhishan Ge\(^2\), Guoyun Lv\(^1\) and Kang Lu\(^1\)

\(^1\) School of Electronic and Information, Northwestern Polytechnical University, No.127, Youyi Xilu, Xi’an, China
\(^2\) Shanghai Electro-Mechanical Engineering Institute, No. 1555, zhongchun Road, Shanghai, China.
Email: lvguoyun101@nwpu.edu.cn

Abstract. Vision-based dynamic multi-objective detection and stable tracking has been a hot topic in recent years. This paper proposed a novel multi-target detection and tracking algorithm based on false-alarms elimination with classification and detection tracking fusion. Firstly, adaptive threshold segmentation and multiple morphological processing were used to finish target detecting, and center-peripheral gray scale difference classification method was proposed to remove false alarm target; secondly, the improved spatial phase correlation algorithm was used to complete the tracking of the target one by one; and thirdly, the target detection and tracking fusion algorithm based on a minimum cost function was also proposed to reduce the false alarm rate of the target. Experiment tests show that compared with the traditional method, the target detection accuracy of the detection and tracking fusion algorithm improves by 7% and the algorithm is able to adaptively adjust the tracking frame size and recapture the target after the target is lost. It can satisfy multiple targets tracking at the same time.

1. Overview
Target detection and tracking based on visible light or infrared video has been a research hotspot in the field of video imaging because of its good concealment, strong anti-interference ability, passive work, and anti-stealth. At present, traditional optical flow methods, such as the detection and tracking technology of moving targets based on the optical flow method proposed by Q Pei, have achieved good results in target detection and tracking [1]. In the field of detection, X Yan et al. proposed the feature detection algorithm of AdaBoost based on extension Haar features [2], solves the problem of high missed detection rate in traditional algorithms. The disadvantage is that the increase in the amount of calculation leads to a decrease in real-time performance. B Yang et al. proposed the SVM detection algorithm based on the ACF feature [3] to solve the problem of slow speed in the detection process. The disadvantage is that the detection effect under complex conditions is not very good. In the field of tracking, based on the phase correlation algorithm, BOLMEDS et al proposed a mean-synthesizing filter [4] to solve the problem of inaccurate positioning, and the disadvantage was that the occlusion did not perform well. DAVID S et al. proposed the filter of the minimum output square error sum [5] to solve the problem of slow speed in the tracking process. The disadvantage is that it is difficult to cope with the deformation of the target.

In the field of detection and tracking fusion, Y Gong proposed a pre-inspection tracking method based on particle filtering [6], which realized the detection and tracking of weak targets. X Zhang et al. proposed detection and tracking algorithm based on multi-feature fusion [7], which solved the problem...
of low detection and tracking accuracy in a specific scenario, but the disadvantage was that the effect in other scenarios was not good.

With the advancement of GPU technology, artificial intelligence has developed rapidly in recent years, and a large number of target detection and tracking algorithms based on deep learning have emerged. The representative of the target detection algorithm is the SSD algorithm proposed by W Liu et al. [8], the real-time performance and precision are both better, but the detection ability for small targets is not very strong. The representative of the target tracking algorithm is MDNET proposed by Z Zhang [9], which solves the problem that the target tracking accuracy is not high, and the disadvantage is that the real-time performance is relatively poor.

In general, the deep learning algorithm has a large amount of computation and big data sets, which is difficult to be applied to embedded devices. Traditional methods are difficult to satisfy the problem of stable anti-jamming tracking. This paper proposes a detection algorithm based on adaptive threshold and classification to false detection and a tracking algorithm of improved related spatial phase, and proposes detection and tracking fusion algorithm based on a cost function.

2. Target Detection Based on Adaptive Threshold and Classification

2.1. Adaptive Edge Detection and Threshold Segmentation

In order to eliminate the effects of noise and interference in the image, the image can be edge-detected by the local standard deviation operator of the image, and then adaptive thresholding is used for binary segmentation.

Image local standard deviation operator, which reflects the contrast change of the gray value of each pixel in a local area of an image. Suppose \( f(x, y) \) is the gray value of the point \((x, y)\) of the \( M \times N \) image, \( w \) is a \( l \times l \) window centered at the point \((x, y)\), \( L \) is odd and \( l \geq L \). \( w \) is a local area of the image, and its local standard deviation is calculated as:

\[
Dev_g = s \sum_{m=-\frac{l-1}{2}}^{\frac{l-1}{2}} \sum_{n=-\frac{l-1}{2}}^{\frac{l-1}{2}} (x_{mn} - \bar{x})^2
\]

In the formula, \( 0 \leq i \leq M - 1, 0 \leq j \leq N - 1 \). \( x \) is the average gray value of all points in the window \( w \), it is defined as follows:

\[
\bar{x} = \frac{1}{l^2} \sum_{m=-\frac{l-1}{2}}^{\frac{l-1}{2}} \sum_{n=-\frac{l-1}{2}}^{\frac{l-1}{2}} x_{mn}
\]

The local standard deviation is affected by the size of the window in equations (1) and (2). To avoid oversmoothing, large values cannot be taken. This article selects the \( 3 \times 3 \) window to calculate the local standard deviation. The output edge image is:

\[
C_{out} = Th(Dev_g(F))
\]

The \( Th(\cdot) \) here takes a threshold calculation.

Based on the Niblack binarization algorithm, this paper proposes an improved adaptive threshold algorithm. The threshold for each pixel is calculated from the dynamically varying standard deviation \( R \). At the same time, the local mean is multiplied by \( R \) and a fixed value \( k \), which can adaptively amplify the effect of the standard deviation.

\[
T(x, y) = m(x, y) \cdot \left[ 1 + k \cdot \frac{s(x, y)}{R} - 1 \right]
\]
Formula (4) is an image binarization formula. M(x,y) and s(x,y) are the mean and standard deviation of each point in the image in its r*r field. R is the standard deviation of the dynamic range; the parameter k takes a positive value.

2.2. Multi-Level Morphology Operation
Because of the influence of noise and other factors, the detected target contour cannot form a closed domain, and multi-level morphological processing is adopted to obtain the target detection result of the closed connected domain.

2.3. Classification to Remove False Alarm Judgment Criteria
With adaptive threshold detection and morphological operations, possible suspect targets will all be detected, but most of them may be false alarms due to the background environment. In this paper, we propose a decision rule based on the change of gray difference around the center.

Since the detected target is located at the center of the target frame, and its gray value has a large difference from the surrounding background, the gray value in the virtual frame does not have such a specific structural distribution. Therefore, this article defines the Central-around Gray difference (CG) feature to represent this structural difference: The elements of the position (i,j) in the matrix L can be defined as:

$$CG(I) = \sum_{c}^{8} ||I_{c} - I_{i,j}||$$

Wherein, I is the detected grayscale image of the target block, its size is normalized to 15*15, and is divided into 9 equal parts, wherein $I_{c}$ is the middle block. $I_{i,j}$ is represented as a surrounding area block of $I_{c}$, $\sum_{c}^{8} ||I_{c} - I_{i,j}||$ indicating that the value added after the corresponding pixel is subtracted.

Through experiments, the detected target box was manually calibrated, and 1893 positive samples and 485 negative samples were selected for CG eigenvalue calculation and statistics. It can be found that when the threshold of taking CG is -2, a portion of negative samples can be effectively eliminated, while most of the positive samples are retained.

3. Based on Improved Spatial Phase-Related Target Tracking

3.1. Pretreatment Process
Target tracking is to determine the position of the target in the tracking search area, and the target's movement will cause the target background to change. Therefore, the window function (Hamming window) filtering needs to be performed on the tracking search area to reduce the influence of the change of the surrounding environment on the correlation filtering. The size of the target tracking search area is updated in real time based on the position and size of the target in the previous frame of the image.

The initialization of the tracking method. In the first frame, the target is manually calibrated using dashed lines. The target state is expressed as:

$$t = [C_x, C_y, M, N]$$

In the formula, $C_x$ and $C_y$ are the center coordinates, M and N are the width and height of the target respectively. At the same time, the search window is used to locate the target and detect the target displacement. The search window is expanded based on the target center coordinate:

$$s = [C_x, C_y, \eta M, \eta N]$$

The larger the value $\eta$, the larger the search window. In order to effectively eliminate the boundary effect and obtain periodic images, the window function $\kappa$ is used to multiply the original image:
In the formula, $I$ is the original image, $x$ is the preprocessed image. After window function filtering, the edge details of the tracking search area can be removed, and the effect of the background change on the algorithm is eliminated. After the filtered image $x$ is Fourier transformed, the filter is:

$$ H = G / X $$

In the formula, $G$ is the spatially distributed spectrum.

### 3.2. Tracking Process

In this paper, an improved spatial phase correlation algorithm is used to determine the position of the target in the current image. If the target spatial offset is $(\Delta m, \Delta n)$, the current frame image $y$ can be written as:

$$ y(m', n') = x(m + \Delta m, n + \Delta n) $$

In the formula, $x$ is the target template; the spatial domain shift $(\Delta m, \Delta n)$ will change the frequency domain phase spectrum of the current frame image, i.e:

$$ Y = X \cdot e^{-(2\pi i / M\Delta m + 2\pi i / N\Delta n)} $$

The formula can be written as a combination of amplitude spectrum and phase spectrum, i.e:

$$ R = \left| \frac{Y}{X} \right| \cdot \left| G e^{-(2\pi / M\Delta m + 2\pi / N\Delta n) \angle G} \right| = \left| G e^{-(2\pi / M\Delta m + 2\pi / N\Delta n) \angle G} \right| $$

In the formula, $\angle$ is the phase, $G$ is the spectrum of the spatial response distribution. Therefore, the spatial displacement of the target is only related to the phase difference and has nothing to do with the amplitude. In fact, the target appearance model cannot always remain unchanged, these changes have little effect on the phase, but the influence on the amplitude result in $\|Y\|$ not equal to $\|X\|$. If the response function is a $\sigma$ function, the performance of the response will be reduced by $\|Y\| / \|X\|$. The spatial response distribution spectrum adopts the spatial distribution proposed in this paper, which can eliminate the influence of differences in the appearance of the model between images. Equation (13) reflects that the response can be decomposed into a spatial response distribution spectrum and the phase difference of the two frames.

### 3.3. Update Process

The tracking performance of the algorithm has a great relationship with the correct template update, so we need to evaluate the results in the tracking to avoid erroneous updating of the template.

Here we define the peak sharpness measure:

$$ PSR = \frac{r_{max} - \mu}{\sigma} $$

Equation (14) shows the relative height between the peak and its surroundings. In equation (14), $\mu$ and $\sigma$ are the mean and standard deviation of the spatial response distribution respectively. Whether the goal applies to template updates can be measured by PSR, with larger values indicating higher confidence in the goal.

### 4. Target to Remove False Alarm Algorithm Combining Detection and Tracking

The tracking and detection fusion module mainly judges the number of real targets in the current frame and their respective current states by comparing the tracking results with the detection results. If
there is a new target, the new target is added using initialization and a new tracker is added; if the target number is not changed, the tracker continues to track the target; if the target disappears, the target state is removed and the corresponding track is removed.

The important performance indicator adopted by the tracking and detection fusion module is the overlap ratio. The overlap ratio is a performance index for evaluating the tracking target area. When estimating the target scale, the evaluation of the overlap rate of the tracking algorithm will also be seriously affected. The overlap rate is defined as follows:

\[
\text{overlap} = \frac{\hat{S} \cap S}{S \cup \hat{S}}
\]

Among them, \(S\) and \(\hat{S}\) respectively represent the estimated state and the real state. If the overlap rate exceeds 50%, the tracking is considered successful; otherwise, the tracking fails.

The algorithm steps with new goals appear as follows:

- Initialize the detector. If the tracker is empty, initialize the tracker.
- If the tracker is not empty, calculate the overlap ratio between the tracker and the detector until the number of tracking targets is equal to detection targets.
- If the detector detects a new target, the tracker is initialized with the detected new target and the overlap between the new tracking target and the detector is calculated.
- It is judged whether it is equal to the target number. If yes, it ends. The algorithm steps with the goal unchanged or the goal reduced are as follows:
  - Determine if the number of trackers and detectors is the same. If not, and the target is detected to be reduced, the disappeared target is removed from the tracker.
  - If the number is the same, determine whether the coverage rate of the tracking result and the detection result exceeds 0.1. If not, use only the tracking result.
  - If the coverage rate exceeds 0.1, fuse; adopt the position of the tracker and the scale of the detector.
  - It is judged whether it is equal to the target number. If yes, it ends.

5. Experiments and Evaluation

5.1. Evaluation Standard

For the real-time performance of the algorithm, this paper uses the average time cost of 100 image sequences to evaluate the performance of the algorithm. The evaluation formula is:

\[
\bar{t} = \frac{\sum t_i}{n}
\]

In the equation, \(t_i\) is the time overhead for the algorithm to process a frame of image, \(n\) is the number of frames of the video sequence used for testing, \(\bar{t}\) is the evaluation coefficient of the algorithm.

For the detection and tracking accuracy of the algorithm, we calculate the difference between the mean value and the real position and size of the target by using the position result \((\Delta x, \Delta y)\) and \((\Delta w, \Delta h)\), and the evaluation formula is:

\[
(\Delta x, \Delta y) = \left(\frac{\sum x_i}{n} - x, \frac{\sum y_i}{n} - y\right)
\]
The detection accuracy of the detection algorithm defines the target detection accuracy. The formula is:

$$\eta = \frac{P}{S + P}$$

In the formula, P is the target number and S is the number of false alarms.

5.2. Experimental Results and Evaluation

It can be seen from the experimental results that this algorithm is superior to the traditional algorithm in terms of real-time performance and the detection and tracking accuracy. The target detection accuracy of detection and tracking fusion improves by 7% comparing with the traditional detection algorithm. Besides, the detection and tracking fusion algorithm is able to adaptively adjust the tracking frame size and recapture the target after the target is lost.

| Algorithms                        | $t_i$   | $(\Delta x, \Delta y)$ | $\eta$ | rechange | recapture |
|----------------------------------|---------|------------------------|--------|----------|-----------|
| Traditional edge detection and classification | 0.895s  | (3,6)                  | 0.85   | N        | N         |
| Adaptive threshold edge detection and classification | 0.672s  | (2,4)                  | 0.90   | N        | N         |
| Traditional spatial phase correlation | 0.069s  | (4,5)                  | N      | Y        | N         |
| Improved spatial phase correlation | 0.021s  | (3,2)                  | Y      | N        | Y         |
| Detection and tracking fusion     | 0.680s  | (2,3)                  | 0.92   | Y        | Y         |

6. Conclusion

This paper proposes a novel multi-target detection and tracking algorithm based on false-alarms elimination with classification and detection tracking fusion. Based on the traditional algorithms, we propose adaptive threshold edge detection and classification algorithm, improved spatial phase correlation algorithm and detection tracking fusion algorithm. Results show that these algorithms are superior to the traditional algorithms in terms of real-time performance, the detection and tracking accuracy and so on.

7. Acknowledgement

This work was supported in part by China shanghai space fund under project SAST2017-004, we would like to thank all master candidates in room 302 for data collection and annotation.

8. References

[1] Pei Q 2009 Detection and Tracking of Moving Objects Based on Optical Flow North China University of Technology

[2] Yan X, Ren W and Ma J 2015 Ada Boost face detection algorithm based on extended Haar feature Computer Systems & Applications 9 pp 152-5.

[3] Yang B, Yan J, et al. 2014 Aggregate channel features for multi-view face detection pp 1-8.
[4] BOLME D S, DRAPER B A and BEVERIDGE J R 2009 Average of synthetic exact filters *IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA, IEEE* pp 2015-22
[5] DAVID S, BOLME J, ROSS B B A, et al 2014 Visual object tracking using adaptive correlation filters *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 no.1 pp 996-1010
[6] Gong Y, Yang H, Hu W, et al. 2007 Detection algorithm for weak target before detection based on particle filter *Systems Engineering and Electronics* 29 no.12 pp 2143-48
[7] Zhang X, Li F, Qin X and Zhang X 2018 An approach to pedestrian detection and tracking based on multi-feature fusion *Software Guide* pp 1-5
[8] Liu W, Anguelov D, Erhan D, et al. 2015 SSD: Single Shot MultiBox Detector pp 21-37
[9] Xing F, et al. 2017 MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network pp 3549-57
[10] Cuchiara R, Piccardi M and Prati A 2003 Detecting moving objects, ghosts, and shadows in video streams *IEEE Transaction on Pattern Analysis and Machine Intelligence* 25 pp 1337-42
[11] Lin R, Du Z and Sun L 2009 Moving Object Tracking based on Mobile Robot Vision *International Conference on Mechatronics and Automation* pp 3625-30.
[12] Grzegorz M. Wojcik, Wieslaw A. Kaminski 2008 Liquid State Machine Built of Hodgkin–Huxley Neurons and Pattern Recognition *Neurocomputing* 58 no.1 pp 245-51.
[13] Niluthpol Chowdhury Mithun, Nafi Ur Rashid and S M Mahbubur Rahman 2012 Detection and classification of vehicles from video using multiple time-spatial images *IEEE Transactions on Intelligent Transportation Systems* 13 pp 1215-25.