Research on Target Tracking Algorithm Based on context Multi-feature Fusion

Yurong Zhao1*, Yao Ling1, Peng Xia1, Qianqian Xu2
1 Anhui Xinhua University, Hefei, Anhui, 230000, China
2 School of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang, 310012, China

*Corresponding author’s e-mail: yb@axhu.edu.cn, 250915025@qq.com;

Abstract. In traditional target tracking algorithms, cosine window is often used to suppress the boundary effect and expand the search range when the target is located, which will reduce the available background information. Besides, when a single feature is used, the representation of the target appearance is not strong. When the challenge factors such as background clutter and occlusion occur to the target, the target will drift, which will reduce the accuracy of target tracking and even lead to tracking failure. To solve this problem, this paper proposes a context-based target tracking algorithm that fuses multiple features, using directional gradient histogram features and color name features as input features. To solve the occlusion problem, the occlusion mechanism of average peak correlation energy and maximum response fraction value is designed to judge the updating of the model. When both values are higher than a certain historical mean, the model is updated; Otherwise, stop updating. Finally, experiments show that the tracking algorithm proposed in this paper is more robust than some traditional correlation filtering algorithms.

1. Introduction
Target tracking[1] is widely used in computer vision, such as video surveillance, human-computer interaction, robots, medical diagnosis[2]. However, in practical application, the tracking accuracy and robustness are still relatively poor under such complex scenes as illumination change, scale change, occlusion, background clutter, motion blur and so on.

The two main challenges of target tracking are target representation and detection sampling. Single-target tracking algorithms usually use the traditional manual Histogram of Gradient (HOG)Feature, Color-Naming feature and learning feature to distinguish target representation. Recently, some researchers have trained depth characteristics on large data sets such as Image Net, used to indicate the target being tracked. Detection sampling is a trade off between computing time and accurate scanning of the target region of interest.

At present, target tracking based on correlation filter has attracted a lot of attention due to its high efficiency calculation using fast Fourier transform. The idea is to return the cyclic version of all input features to the target Gaussian function, so there is no need for a hard threshold sample of the target appearance. Bolme et al. Learning the minimum output of the squared error filter on the luminance channel and for fast tracking, several extensions have been proposed to significantly improve the tracking accuracy, including nucleation dependent filters, multidimensional characteristics, Context learning, scale estimation, scale adaptive tracking based on multi-feature fusion and complementary
features to learn real-time tracking\cite{3}.

However, the main disadvantage of correlation filter trackers is the boundary effect due to cyclic displacement. In addition, in order to limit the target drift, the target search area contains only a small local neighborhood. The cosine window is used to suppress the boundary effect, and the search range is reduced effectively. Therefore, the correlation filter tracker has very limited context information, and it is easy to drift when the target is moving quickly, occlude or the background is chaotic. In view of the above challenges, we design the framework of target tracking based on context more features fusion, first of all, using the traditional feature extraction method, the extraction of HOG features and CN through simple fusion of vectors, and second, the up and down or so four blocks to video frame target feature extraction, using correlation and convolution filter learning to generate their own response diagram, get the maximal response of the estimates of the target location, and then by judging the average peak energy (APEC) Whether the values of the two are higher than a certain historical mean determines whether the filter model is updated. Adopt the object tracking benchmark (OTB-100) Video sequences with multiple challenge factors were tested and compared with three mainstream tracking algorithms based on correlation filtering. The experimental results fully proved that the proposed algorithm was superior to other algorithms in robustness and tracking accuracy.

2. Nuclear correlation filter target tracking

Correlation filter tracking is based on discriminant target tracking. The purpose is to train a correlation filter to filter the target image block, transform the calculation into frequency domain, reduce the complexity of calculation, and estimate the location of the target by the location of the maximum response value.

Set the training sample set as \((x_i, y_i)\):

\[
\min_h \left\| B_0 h - y \right\|_2^2 + \lambda \left\| h \right\|_2^2
\]  

(1)

Here, \(h\) represents the correlation filter template, all cyclically shifted sample sets of the target image block with a matrix \(B_0 = [x_1, x_2, \ldots, x_n]\), \(y\) is the sample label, \(\lambda\) is the regularization coefficient to prevent over fitting.

Under the linear condition, the solution of the correlation filter \(h\) can be obtained by using the property of cyclic matrix:

\[
H = \frac{B_0^* \odot Y}{B_0^* \odot B_0 + \lambda_i}
\]  

(2)

\(H\) is the Fourier transform of \(h\), \(B_0\) is the Fourier transform of \(b_0\), and \(B_0^*\) is the conjugate of \(B_0\).

Detection formula, in the next frame, filter \(h\) convolved with the image block \(z\) (search window), where \(F(\ )\) represents the Fourier transform, \(F^{-1}(\ )\) represents the inverse Fourier transform, \(Z\) represents the Fourier transform of \(z\). The location of the maximum response is the target location within the search window. The original domain detection formula is given by the following formula:

\[
F(h) = Z \odot H
\]  

(3)

At this point, the maximum response value is \(F^{-1}\left(F(h)\right)\), which is the location of the target.

Ridge regression in nuclear space:

\[
\hat{\alpha} = \frac{Y}{B_0^* \odot B_0 + \lambda_i}
\]  

(4)

\(\hat{\alpha}\) is the Fourier transform of \(\alpha\).

For the Gaussian kernel, then:

\[
K^{\alpha} = \exp\left(\frac{1}{\sigma^2}\left(\left\|\alpha\right\|_2^2 + \left\|\beta\right\|_2^2 - 2\left(F^{-1}(\alpha^\prime \odot X)^\prime\right)\right)^2\right)
\]  

(5)

The dual variables \(\alpha\) can be used directly for detection by representing the original variable.
In the process of target tracking, the appearance of the target will change. In order to continuously track the target, an online update filter is required. When the target tracking is carried out on the frame of frame $t$, the target model image learned is defined as $\overline{B}'$, and the update formula of the relevant filter model $h$ is:

$$
\begin{align*}
\overline{B}' &= (1-\eta)\overline{B}'^{-1} + \eta \overline{B}' \\
\overline{H}' &= (1-\eta)\overline{H}'^{-1} + \eta \overline{H}'
\end{align*}
$$

(6)

Where, $\eta$ is the learning rate.

3. Context multi-feature fusion target tracking

The features of traditional feature orientation gradient histogram (HOG) and color name (CN) are introduced. Secondly, the judgment of occlusion mechanism is introduced in detail. Then, how to add context information in the correlation filter tracking framework is introduced. In order to avoid the drift of the target as much as possible, the target appearance of the first frame is annotated.

3.1. Feature Introduction

Directional gradient histogram feature is mainly used to describe the local texture image, high calculation efficiency, popular in the field of vision. Since the HOG feature is extracted in the image unit (cell), it has a good feature of non-deformation.

The color name or color attribute feature is a perspective space used to describe the color of the language color label. Color label space distance is better than RGB. In order to achieve promising results in visual fields such as target recognition, target detection and action recognition, we use the mapping method described in the above to transform RGB space into the color name space represented by 11 dimensional color. The color name provides the perception of the target color and contains important information related to the target.

HOG emphasizes the target location information, while the color name focuses on the color information. In the process of target tracking, the color feature (CN) can well deal with the deformation and motion blur of the target. However, when there is a strong light change, the characterization ability of the color feature becomes weaker. The HOG feature can well adapt to the change of care and carry out accurate tracking.

$$
K^v = \exp\left(\frac{1}{\sigma^2} \left[ \|x\|^2 + \|x - 2 \left( F^{-1}(X_c \odot X_c) \right) \|^2 \right] \right)
$$

(7)

In the formula, $X_c$ refers to the feature of the superposition of vectors.

3.2. Judgment of occlusion mechanism

At present, most tracking algorithms update strategies are relatively simple. When updating the tracking model of each frame of video, the accuracy of tracking results is not taken into account. When the target is severely blocked or the target tracking results are inaccurate, it is easy for target tracking to fail.

(a) No occlusion

(b) Occlusion

Figure 1. Trace results response diagram
The confidence of the trace results can be reflected by the peaks on the response graph. For example, when the target of the current frame is accurately detected, only one peak appears in the response graph, and the other positions are smoother, as shown in Figure 1(a), which is the ideal response graph. If the target is severely obscured, continue to update the tracking model with uncertain training samples, and the tracking model will be contaminated, resulting in target tracking failure, as shown in Figure 1(b). Based on the above situation, this paper uses an occlusion mechanism to judge whether to update the model. First, the maximum response score after feature fusion is calculated, which is defined as:

$$F_{\text{max}} = \max F(s, y : w)$$

Then, calculate the Average peak correlation energy ratio, and define it as:

$$\text{APCE} = \frac{\sum_{w} (F_{\text{max}} - F_{w})}{\text{mean} \left( \sum_{w} (F_{\text{max}} - F_{w}) \right)}$$

Where, $F_{\text{max}}$, $F_{\text{min}}$, and $F_{w,h}$ represent the maximum, minimum, and the elements of row $w$ and column $h$ in $F(s, y : w)$ respectively. $\text{APCE}$ measures the degree of volatility in response graphs. Within the detected area, when the target is obvious, the response peak value will become sharp and the noise will decrease. only one peak value will appear in the response graph. At the same time, The value of $\text{APCE}$ will be bigger; Conversely, when the target is obscured or lost, the value of $\text{APCE}$ will decreases.

When a tracking result with high confidence appears, the parameters $\text{APCE}$ and $F_{\text{max}}$ of the target in the image of this frame should be greater than the historical mean, which satisfies formula (10).

$$\text{APCE} > a \times \text{APCE}_{\text{mean}}$$

$$\& \& F_{\text{max}} > b \times F_{\text{max}} - \text{mod el}$$

When the target is blocked, its response graph fluctuates greatly with multiple peaks, as shown in Figure 1 (b). At this time, the current frame stops updating the model to ensure that the target can still be tracked in the next frame of the video. On the contrary, if the model continues to be updated, the occlusion will affect the tracking result and lead to the tracking failure.

3.3. Target tracking
The tracking performance is affected by the surrounding environment, occlusion and other factors. This paper proposes to add context information to the target tracking framework of the filter. In order to make effective use of background information, a block background image is selected around the target and denoted as $b_i, i \in (1, n)$. By cyclic displacement of $b_i$, $B_i$ can be obtained. The objective function is:

$$\min_h = \left\| B_i h - y \right\|_2^2 + \lambda \| h \|_2^2 + \lambda \sum_{i=1}^{k} \| B_i h \|_2^2$$

Can be seen from the formula (11) compared with the traditional correlation filtering tracking a penalty term $\lambda \sum_{i=1}^{k} \| B_i h \|_2^2$. In the tracking algorithm in this paper, $B_i$ context region is introduced as a negative sample, and the intermediate target sample $B_0$ as a positive sample. In the calculation process, the value of training template $h$ and the correlation response of negative sample $B_i$ are minimized as far as possible, and the closed solution is obtained. The solution results of single-channel characteristics are as follows:

$$H = \frac{X_i \odot Y}{X_i \odot X_i + \lambda_i + \lambda \sum X_i \odot X_i}$$

Meanwhile, the solution of the maximum correlation response is the same as that of the traditional
correlation filtering algorithm, which can be obtained by formula (2).

The ridge regression equation of multi-channel characteristics can be set as follows:

$$f(h) = \|Ph - \bar{y}\|^2 + \lambda \|h\|^2$$  \hspace{1cm} (13)

The magnitude of $\tilde{P}$ is $(n+1)^{k^m}$, and $m$ represents the number of characteristic channels. The solution results are as follows:

$$h = (\tilde{P}^T\tilde{P} + \lambda I)^{-1}\tilde{P}^T\bar{y}$$  \hspace{1cm} (14)

Where $h$ is the correlation filter of the learning context.

The formula of single channel feature detection is consistent with that of target detection. Target to keep out problem, the algorithm introduces a new model update strategy, using two relevant judgment on the basis of the average peak energy and maximum response points. When the APCE and $F_{\text{max}}$ criteria of an image to be searched are both higher than a certain historical mean, this frame will be considered to have high confidence. At this time, the filter model will be updated, thus reducing the pollution of the filter model and improving the accuracy of target tracking.

3.4. Algorithm Steps

1) Filter learning stage: HOG and CN features are extracted from the target image block $B_0$ and the upper and lower background image block $B_i$ in the target frame $t$ by estimating the target position.

2) The HOG and CN features extracted are added and fused through a simple vector, and the fused features are then transformed into fast Fourier transform (FFT) to train the correlation filter in the frequency domain.

3) Detection stage: at the $t+1$ frame, HOG and CN features were extracted and fused in the search area, fast Fourier transform (FFT) was performed, and the filter model $t+1$ was used for point-multiplication in the frequency domain. Then the inverse Fourier Transform (IFFT) was used to obtain the response graph, and the target position was determined according to the position of the response graph.

4) Judge whether to update the filter model according to Formula (10), that is, update the model when the average peak correlation energy (APCE) and the maximum response value fraction ($F_{\text{max}}$) are both higher than a certain historical mean; otherwise, do not update the model, and proceed to the next frame.

The algorithm flow chart is shown in Figure 2.

![Algorithm flow chart](image)

4. Analysis of experimental results

In this paper, the experimental environment based on Matlab2016, Windows 7 system using benchmark target tracking (OTB-100) to evaluate the performance of the algorithm, part of the video
sequence with illumination change, plane internal/external rotation, fast moving, motion blur, background interference, shade, dimension change and the deformation challenge factors such as rigidity, and comparing with the mainstream of the tracking algorithm. Mainly use the Distance accuracy (short Precision, DP), overlapping degree (Overlap Precision, OP), the average Frame rate (Frame Per Second, and FPS), evaluated the three indexes including DP describes the tracking algorithm to estimate the center of the target location and artificial mark the center of the target, the Distance of the two is less than the given threshold video frames, percentage of threshold is set to 20 pixels, to evaluate the robustness of the algorithm; OP refers to the percentage of the number of frames whose score is greater than a certain threshold in the total number of frames tracked, and evaluates the accuracy of the algorithm. According to PASCAL evaluation index, the threshold of overlapping rate of this text is 0.5, following the protocol, the same parameter values are used for all video sequences and all sensitivity analysis, and the tracker is implemented on Intel I5-4770 in Matlab.

4.1. Qualitative analysis

Figure 3 shows the algorithm in this paper by conducting tracking tests on two video sequences (a) Liquor_1 and (b) Jogging_1. In order to visually demonstrate the advantages of the proposed algorithm, it is compared with traditional KCF algorithm, feature-based correlation filter tracking algorithm SAMF algorithm and Staple algorithm.

4 representative tracking results were selected for each video for comparison. In the figure, different wire frames are used to represent the four algorithms to mark the location of the target during tracking. In video sequence (A) Liquor_1, in the process of tracking, the target can be accurately and effectively tracked when there are challenge factors such as motion blur and interference with similar objects. In the video sequence (b) Jogging_1, when the target is completely blocked, the algorithm in this paper adopts the judgment of APCE blocking mechanism to update the model adaptively, avoid the pollution of the model, accurately track the target in the subsequent frames, and show a strong tracking robustness. In general, compared with traditional KCF algorithm, SAMF algorithm based on feature fusion and Staple algorithm, this algorithm has improved its target tracking performance to a certain extent.

4.2. Quantitative analysis

Occlusion, rapid movement, background blur, rotation, scale change and other challenging factors were selected for the video sequence in the experiment. Table 1 shows the average precision distance (DP) and average overlap accuracy (OP) of the proposed algorithm and THE SAMF, Staple and KCF
algorithms on 21 video sequences. It can be seen from Table 1 that the tracking accuracy and success rate of the algorithm in this paper are 88.3% and 79.7% respectively. Compared with SAMF algorithm, DP has been improved by 9.6% and OP by 3%, both showing good tracking effect.

| The evaluation index | SAMF | Staple | KCF | OUR |
|----------------------|------|--------|-----|-----|
| Speed (FPS)          | 82   | 87     | 128 | 78  |
| Business, DP (%)     | 77.7 | 75.4   | 62.8| 88.3|
| Business, OP (%)     | 76.7 | 69.8   | 60.0| 79.7|

Figure 4 shows the accuracy and tracking success rates of the above four algorithms in OTB-100. The algorithm in this paper is named OUR, and its tracking accuracy reaches 0.883, which is much higher than SAMF algorithm, Staple algorithm and KCF algorithm, and the corresponding tracking accuracy is 0.77, 0.754 and 0.628, respectively. The success rate of OUR algorithm is 0.797, slightly higher than that of SAMF algorithm, and much higher than that of KCF algorithm (0.60). It is clear from Figure 4 that OUR algorithm achieves good results in terms of both tracking accuracy and robustness.

Figure 4 shows the accuracy and tracking success rates of the above four algorithms in OTB-100. The algorithm in this paper is named OUR, and its tracking accuracy reaches 0.883, which is much higher than SAMF algorithm, Staple algorithm and KCF algorithm, and the corresponding tracking accuracy is 0.77, 0.754 and 0.628, respectively. The success rate of OUR algorithm is 0.797, slightly higher than that of SAMF algorithm, and much higher than that of KCF algorithm (0.60). It is clear from Figure 4 that OUR algorithm achieves good results in terms of both tracking accuracy and robustness.

Figure 4 shows the accuracy and tracking success rates of the above four algorithms in OTB-100. The algorithm in this paper is named OUR, and its tracking accuracy reaches 0.883, which is much higher than SAMF algorithm, Staple algorithm and KCF algorithm, and the corresponding tracking accuracy is 0.77, 0.754 and 0.628, respectively. The success rate of OUR algorithm is 0.797, slightly higher than that of SAMF algorithm, and much higher than that of KCF algorithm (0.60). It is clear from Figure 4 that OUR algorithm achieves good results in terms of both tracking accuracy and robustness.

5. Conclusion

In this paper, a target tracking framework based on context multi-feature fusion is proposed to incorporate context into the training phase of correlation filter at a lower computational cost. In this paper, the direction gradient histogram (HOG) feature is fused with the vector with simple color name (CN) feature as the input feature, finally designs the related energy (APCE) and maximum average peak response mechanism of shade of the score value judgment model update, so as to improve the tracking performance of algorithm, the experimental results show that this algorithm can use shade in
target, background clutter, fast moving and scale change under complex scene, several related than traditional filtering algorithm has better accuracy and robustness. However, there are still some shortcomings in this paper. The feature extraction model of target image block in video frames can be improved, and the feature fusion method can be improved to improve the real-time and accuracy of tracking algorithm.

Figure 5. The accuracy and success rates of tracking test benchmarks in complex scenarios

Acknowledgments
This work was financially supported by the Scientific Research Project of Anhui Xinhua University (2019zr012); (2019zr002); (2019zr003); (2018jpckx31) and Anhui Quality Engineering Project (2019mooc348).

References
[1] GANAPATI V, PLAGEMANN KOLLK, ER D, et al. Real Time Motion Capture Using a Single time-of-flight Camera//Proc of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE, 2010:755-762.
[2] Bonin-font F, ORTIZ A, OLIVER G. Visual Navigation for Mobile Robots: A Survey. Journal of Intelligent and Robotic Systems, 2008, 53(3): 263-296.
[3] REVELL J, MIRMEHDI M, MCNALLY D. Computer Vision Elastography: Speckle Adaptive Motion Estimation for Elastography Using Ultrasound. IEEE Transactions on Medical Imaging, 2005, 24(6):55-766.