Target detection and tracking based on improved HOG-color feature fusion

Tianle Dai, Ziqiang Hao*, Zhongyuan Wang, Weida Zhan and Yanfeng Tang

School of electronic information engineering, Changchun University of Science and Technology, Changchun, China

*Corresponding author e-mail: 7663410@qq.com

Abstract. In the process of constructing the target appearance model, the traditional HOG feature does not take into account the relationship between adjacent regions when the image is segmented in the feature extraction process, resulting in the regional aliasing effect, and the single feature is inefficient in target detection and tracking. An improved HOG-color feature fusion method for target detection and tracking is proposed. Firstly, HOG feature and color feature were extracted from the target samples respectively. During the HOG feature extraction process, trilinear interpolation was used to eliminate regional aliasing effect. Secondly, Bhattacharyya distance is used to select the HOG feature after interpolation, and the appropriate feature is selected as the HOG feature. Then combine the HOG feature after selection with the color feature; Finally, the filter is obtained by learning the kernel correlation filter, and the image is correlated detected with the filter, and the response output is obtained. The experimental results show that this method is superior to the traditional HOG-color feature fusion target detection and tracking method in both tracking speed and tracking efficiency.

1. Introduction

With the rapid development of image processing and machine learning technology, object detection and tracking technology have attracted more and more researchers' attention. Target detection refers to the process of extracting changing regions from background images by means of algorithm from video image sequence, which is the basis of target identification, classification and target tracking. Target tracking refers to the process of using computer technology to find the position of an independent moving target in the video sequence and its complete motion trajectory [1].

Henriques et al. proposed a KCF tracking algorithm based on circulation structure, and introduced the kernel method into MOSSE (Minimum Output Sum of Squared Error filter), achieving better tracking performance [2]. Danelljan et al. adopted more complicated HOG features on the basis of MOSSE and introduced an estimation strategy for target scale, which improved the tracking accuracy significantly, but the speed of the algorithm needed to be improved [3]. Martin D et al. introduced adaptive scale tracking and designed DSST(Discriminative Scale Space Tracking) algorithm. they used two filters to train position and scale model respectively, and achieved good results in moving variable scale target tracking [4]. Li et al. Put forward a method to fuse the HOG features and color features on the basis of KCF, which improves the description of the object appearance by enhancing the representation dimension of the object. This method has achieved better tracking performance, but because of the HOG feature extraction process does not take into account the relevance between the
connected block area, leading to aliasing effect produced in the process of feature extraction affect the efficiency of target detection and tracking [5]. In this paper, an improved target detection and tracking algorithm based on HOG-color feature fusion is proposed, which can effectively remove the region aliasing effect of traditional HOG features in the process of feature extraction. The performance of target tracking is improved obviously.

2. Improved HOG feature extraction

In the traditional HOG feature extraction process, it is not considered that when the image is segmented, the boundary pixels have a certain correlation with the connected region. If only the current region is calculated and the relationship with the connected region is completely ignored, the regional aliasing effect will be generated. This aliasing effect will cause the mutation of feature vector during feature extraction. Trilinear interpolation [6] was used to modify the histogram of direction gradient vector of each block. The gradient amplitude of a certain pixel on the boundary is accumulated to the corresponding bin with different weights.

Suppose that the center pixels of the four cell units in a block are respectively \((x_1, y_1), (x_2, y_1), (x_2, y_2), (x_1, y_2)\). The histogram of the gradient direction of the four cell units are respectively \(h(x_1, y_1, \theta), h(x_2, y_1, \theta), h(x_2, y_2, \theta), h(x_1, y_2, \theta)\), \(\theta \in \{k \times \pi/9, k \in 1, \ldots, 9\}\).

Taking the center of the angle interval as the center value of each bin in the histogram, the gradient amplitude of the pixel to be processed is weighted and accumulated to the two bins by the linear relationship between the two adjacent bins. Assuming that the pixel with the gradient direction of 25° is to be processed, it is obvious that 25 is closest to the histogram with the centers of 10 and 30, the gradient amplitude of the point should be weighted and accumulated on these two histograms, the weights are \((25-20)/20 = 0.25\) and \((25-10)/20 = 0.75\), respectively. As shown in Figure 1.

![Figure 1. The diagram of pixels to be processed by linear weighting](image)

The position coordinates and the gradient direction of the processed pixel are interpolated, and the gradient amplitude of the point is weighted to the corresponding histogram of four cell units. And weighted to the two bins adjacent to the \(\theta\). The correction formula of cell gradient direction histogram is as follows [7]:

\[
h(x_1, y_1, \theta_1) = h(x_1, y_1, \theta_1) + G \left(1 - \frac{x - x_1}{b_x}\right) \left(1 - \frac{y - y_1}{b_y}\right) \left(1 - \frac{\theta - \theta_1}{b_\theta}\right) \tag{1}
\]

\[
h(x_1, y_1, \theta_2) = h(x_1, y_1, \theta_2) + G \left(1 - \frac{x - x_1}{b_x}\right) \left(1 - \frac{y - y_1}{b_y}\right) \left(1 - \frac{\theta - \theta_2}{b_\theta}\right) \tag{2}
\]

Where \(x, y\) represents the position of the pixel, \(\theta\) represents the gradient direction of the pixel, and \(G\) represents the gradient amplitude of the pixel to be processed. \(\theta_1, \theta_2\) represents the midpoint coordinates of the two adjacent bin with \(\theta\). \(b_x = b_y = 8\), \(b_\theta = \pi/9\). Similarly, \((x_1, y_2), (x_2, y_1), (x_2, y_2)\) cell units can be obtained by formula (1) and (2). The HOG features extracted after the gradient direction histogram is modified by the above linear interpolation are shown in figure 2.
3. Improved HOG feature selection

Due to the high dimension of the HOG feature extracted, the calculation of the algorithm is large, which leads to poor real-time performance, and easy to cause target loss when tracking the video image. Therefore, the HOG feature extracted is effectively selected to reduce the dimension of the feature. The usual dimension reduction methods include principal component analysis (PCA), sparse coding, linear discriminant analysis (LDA) and so on. All of the above methods can well extract the main information of the features, but the algorithm has a large amount of calculation and cannot meet the real-time requirements of the system. In order to select more expressive features faster, Bhattacharyya distance [8] is adopted as the basis for feature selection, which is often used to distinguish the separability of two discrete probability distributions. When the histogram similarity is compared, the Bhattacharyya distance can get the best effect and satisfy the condition of class separability. Firstly, the average value of the improved HOG features is calculated statistically, then the separability criterion of each feature is calculated separately, and the larger first features are selected. Bhattacharyya distance is defined as:

$$D_B (p, q) = -\ln \int \sqrt{p(s_1)q(s_2)} \, ds$$

(3)

Where p and q are respectively represented by two discrete probability distributions in the same domain, $s_1, s_2$ represent both positive and negative sample categories of the target to be tested. In this paper, INRIA data set [9] was used, record pedestrians as $s_1$ categories and non-pedestrians as $s_2$ categories. In the INRIA data set, calculate the Bhattacharyya distance of each one-dimensional feature in the sample image, as shown in figure 3.
As can be seen from figure 3, the pasteurization distance of channel 1, 5, 6 and 9 is larger than that of other feature channels, that is, feature discrimination is strong. Therefore, in order to reduce feature extraction dimension, as few features as possible can be selected for training to obtain the classifier. Firstly, the classifier is obtained by training \( \text{bin} \in \{1, 5, 9\} \) and \( \text{bin} \in \{1, 5, 6, 9\} \) respectively. Then, according to the test data set, the experimental results are shown in table 1:

| Feature Channel selection | Positive sample | Negative sample | The total detection rate |
|---------------------------|-----------------|-----------------|-------------------------|
| Improved HOG {1,5,9}      | 97.0523%        | 94.5165%        | 96.7938%                |
| Improved HOG {1,5,6,9}    | 97.8579%        | 95.0475%        | 97.0688%                |

From table 1, it can be seen that the detection rate of channel \( \text{bin} \in \{1, 5, 6, 9\} \) is slightly higher than that of channel \( \text{bin} \in \{1, 5, 9\} \), but for the feature dimension, channel \( \text{bin} \in \{1, 5, 6, 9\} \) is much higher than that of channel \( \text{bin} \in \{1, 5, 9\} \). In order to reduce the calculation amount of the algorithm, a slight increase in the detection rate is not considered. Therefore, this paper finally selects three feature channels as the HOG feature of the sample.

4. Color feature extraction

For a \( H \times W \) size rectangular image region, in the process of establishing a color histogram [10], according to the different contributions of pixels in different positions to the color histogram, the closer the pixel is to the center of the region, the greater the weight, and the farther the pixel is to the center of the region, the smaller the weight. Its weight function is defined as:

\[
    w(d) = \begin{cases} 
    1 - \frac{d}{D}, & d < D \\
    0, & d \geq D 
    \end{cases}  \tag{4}
\]

Where, \( d \) represents the distance from any pixel point to the center of the region.

\[
    c(y) = \{c_b(y), b = 1, 2, \cdots, R\} \text{ represent the histogram of color distribution at the center point at } y. \text{ In order to avoid the impact of image scaling, it can be obtained by normalization:}
\]

\[
    c_b(y) = \frac{\sum\limits_{i=1}^{N} w(d) \frac{||y - x_i||}{\sqrt{H^2 + W^2}} \delta(\theta(x_i) - b)}{\sum\limits_{i=1}^{N} w(d) \frac{||y - x_i||}{\sqrt{H^2 + W^2}}} \tag{5}
\]

Where, \( N \) represents the number of pixels in the target region of size \( H \times W \), and \( \delta(\cdot) \) is the Kronecker delta function. When the ith pixel is located in the b subinterval, it is 1; otherwise, it is 0. \( \theta(x_i) \) represents the subinterval index \( b \) of any pixel \( x_i \) in the color histogram.

5. The improved HOG and color feature fusion

The edge and gradient features extracted by HOG feature can well grasp the features of the local shape, but ignore the flat surface information. As a global feature, the color feature can well describe the target surface features corresponding to the image region. Therefore, in this paper, the improved HOG and HSV color features are fused. This method can better describe the characteristics of the target, effectively reduce the dimension of the target features, and greatly reduce the time of feature extraction.

First, the sample image is segmented into cells of equal size. Then extract the color histogram in each cell, find the maximum color histogram as the HSV color feature of the cell, and combine the
features of four adjacent cells into a block feature, so that each block has a 4-dimensional HSV color feature. Finally, the obtained 12-dimensional HOG feature is fused with the obtained 4-dimensional HSV feature vector. At this time, there are 16-dimensional feature vectors in the obtained block. After fusion, the feature is denoted as F, then F can be expressed as:

$$F = \mu \cdot F_{hog} + (1 - \mu) \cdot F_{hsv}$$  

(6)

Where, \(\mu\) represents the fusion coefficient, \(F_{hog}\), \(F_{hsv}\) respectively represents the HOG feature and color feature extracted. By this method can calculate the samples after fusion image feature vector dimension is \(16 \times 7 \times 15 = 3780\), the use of traditional HOG feature extraction of a size of 64×128 images, get the dates for the characteristic dimension of \(36 \times 7 \times 15 = 3780\), compared the improvement method of HOG feature extraction of sample image feature vector dimension is reduced greatly, so this method can effectively reduce the dimension of the extracted features, reduce the time of the feature extraction and improve the efficiency of target detection and tracking.

6. Experimental results and analysis

In terms of target detection, in order to verify the effectiveness of the proposed algorithm, INRIA pedestrian data set was used for experimental verification. The data set contains 2416 training samples of pedestrians (positive) and 1218 training samples of non-pedestrians (negative). It also includes 1126 pedestrian test samples and 453 non-pedestrians test samples. On the experimental data set, the traditional HOG feature [1 1], the improved HOG feature and the algorithm in this paper are verified and compared respectively. The total detection rate and feature extraction time are used to measure the effectiveness of the algorithm. The experimental results are shown in table 2:

Table 2 Detection results of different methods on INRIA pedestrian

| Feature used     | Feature dimension | Feature extraction time | Total detection rate |
|------------------|-------------------|-------------------------|----------------------|
| HOG features [11] | 3780              | 88.075ms                | 88.35%               |
| Improved HOG     | 1260              | 26.875ms                | 90.94%               |
| Proposed         | 1680              | 28.667ms                | 96.93%               |

As can be seen from table 2, compared with the reference [11], the total detection rate is increased by 8.58%, and the feature dimension is reduced by 2100 without affecting the loss of important features, and the corresponding feature extraction time is shortened by about 30%. If only the improved HOG single feature is used for detection, although it has certain advantages over the algorithm in feature dimension and feature extraction time, there is a certain gap in recognition accuracy. Therefore, in terms of overall detection performance, although the improved HOG-color feature fusion algorithm is slightly higher than the improved single HOG feature in feature extraction time, it has an obvious improvement in recognition rate, about 5.99%.

From the above experiments, it can be seen that the algorithm is better in time and accuracy than the traditional algorithm in target detection. In this paper, the above improved and fused features were also tested on the target tracking, mainly compared with the traditional KCF algorithm and the KCF algorithm integrating different features, and the center position error CLE, tracking accuracy DP, success rate OP and algorithm speed FPS were used as evaluation indicators, and the experiment was conducted on the OTB2015 data set. The experimental results are shown in table 3:

Table 3 Tracking performance between the proposed algorithm and the traditional KCF algorithm

| Algorithm             | CLE | DP  | OP  | FPS |
|-----------------------|-----|-----|-----|-----|
| KCF+HOG               | 45.6| 69.7| 61.2| 143 |
| KCF+HOG+HSV           | 36.8| 72.3| 67  | 47  |
| Improved KCF+HOG+HSV  | 34.3| 78.6| 72.5| 79  |
The table 3 shows that compared with the traditional single feature of HOG kernel correlation filter, the improved HOG-color features fusion algorithm has obvious advantages in the tracking accuracy and success rate, the reason is that in target feature extraction combined with color features, making the features extracted by the tracker more expressive, compared with the traditional HOG-color feature kernel correlation filter, on the tracking accuracy and success rate of this algorithm improved slightly, but this algorithm on the tracking speed obvious advantage, because in the process of HOG features of target are extracted, the trilinear interpolation algorithm is adopted to reduce the aliasing effect of the features of connected regions, making the HOG feature extracted more expressive. In addition, features are selected in the feature extraction process, greatly reducing the dimension of the extracted features and significantly improving the tracking speed of the tracker.

7. Conclusion
In this paper, the traditional algorithm of target detection and tracking based on HOG-color feature fusion is improved. On the one hand, the improved method of target detection and tracking based on HOG-color feature fusion overcomes the problem that the traditional HOG feature ignores the relationship between adjacent regions in the process of feature extraction, which leads to the mutation of features in the process of feature extraction, as well as the large dimension and long time of feature extraction. On the other hand, it also changes the traditional single feature in the aspect of target detection and tracking, which is easily affected by background, noise, illumination and other factors, resulting in the low rate of target detection and tracking. Experiments show that this method is superior to the traditional HOG-color feature fusion method in both tracking speed and tracking efficiency.

References
[1] Lu Huchuan, Li Peixia, Wang Dong. Overview of target tracking algorithm [J]. Pattern recognition and artificial intelligence,2018,31(01):61-76.
[2] Henriques J F, Rui C, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[C]. European Conference on Computer Vision, 2012:702-715
[3] Danelljan M, Hager G, Khan F S, et al. Accurate scale estimation for robust visual tracking[C]. British Machine Vision Conference, 2014:65.1-65.11
[4] Smeulder A W M, Chu D M, Cucchiara R, et al. Visual Tracking: An Experimental Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1442-1468
[5] Danelljan M, Khan F S, Felsberg M, et al. Adaptive color attributes for real-time visual tracking[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014:1090-1097
[6] Pang Y, Yuan Y, Li X, et al. Efficient HOG human detection[J]. Signal Processing, 2011, 91(4):773-781.
[7] Huang Dongli, Dai Jianwen, Feng Chao, Zhang Jinli, Yu Shuang, Zhang Fang. Trilinear interpolation algorithm for HOG feature extraction [J]. Computer knowledge and technology,2012,8(31):7548-7551.
[8] Xiong Jianbin, Wang Qinruo, Deng Jiuying, Liu Qi, Ye Baoyu. Face recognition based on the combination of Barclay distance and PCA [J]. Computer engineering and application,2012,48(03):202-204+215.
[9] INRIA Database[OL]. http://pascal.inrialpes.fr/data/human/
[10] Sapp B, Toshev A, Taskar B. Cascaded Models for Articulated Pose Estimation // Proc of the 11th European Conference on Computer Vision. Heraklion, Greece, 2010: 406 – 420
[11] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Computer Vision and Pattern Recognition, 2005 :886-893