Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Optimization Techniques for Target Tracking in Video Image Signal Processing Based on Multiscale Convolutional Neural Networks

Xinming Fan[1,✉]

[1] *School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu, 224051, China*

ABSTRACT

Target tracking is a fundamental task in the field of computer vision, which has a wide range of applications in real-life video image signal processing. This paper proposes target tracking optimization technique based on the principle of multi-scale convolutional neural network and multi-target tracking algorithm. The basic structure is designed using VGG16 network, the ROI align method is used to reduce the number of features for feature fusion, and the improved Hungarian algorithm is adopted to associate the fused features and obtain the target tracking results. In the tracking performance experiments, the target tracking optimization technique in this paper is more discriminative in terms of extracted features, and also has higher tracking results under challenging factors such as background clustering (BC), scale variation (SV), and out-of-view (OV). As for the target tracking experiments on mobile network video images, the average tracking accuracy and average tracking success rate of this paper's method are 97.89% and 96.02%, which are better than DS_v2 and FFT16, and the average error between the target tracking results and the target's actual motion trajectory is 4.12mm, while possessing the smallest error amplitude.

Keywords: multi-scale convolutional neural network; VGG16 network; ROI align; Hungarian algorithm; target tracking

1. Introduction

Video information processing technology refers to a series of technologies for processing and analyzing video signals, and its basic principle involves a number of fields such as digital signal processing, image processing, and computer vision. Video information processing technology has a wide range of applications in many fields such as security monitoring, intelligent transportation, medical diagnosis, entertainment media and so on [1-4]. In video information processing technology, image processing is a core link, including image enhancement, image denoising, image segmentation,

✉ Corresponding author.
E-mail address: fwxhfwj@sina.com (X. Fan).
Received 25 February 2024; Revised 10 May 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-212

feature extraction and other steps. Image enhancement refers to a series of algorithms to improve the visual effect of the image, so that it is clearer and easier to analyze [5-8]. Image denoising removes random noise from the image through filtering algorithms to improve the quality of the image. Image segmentation is to divide the image into multiple regions for subsequent feature extraction and analysis. Feature extraction is to extract information from the image that is useful for analysis, such as edges, corners, colors, etc. [9-12].

Target tracking in video image processing is a core task, which aims to accurately recognize and track moving targets in video. This technique has a wide range of applications in many fields, such as video surveillance, intelligent transportation systems, robot navigation, etc. [13-15]. The study of target tracking technology involves several key issues, which include target detection, target tracking algorithms, and tracking stability. Target detection is the first problem to be solved, which aims to accurately extract the target in the image from the background [16-17].

Literature [18] emphasizes on target tracking in video surveillance. The advantages and disadvantages of tracking strategies such as area based and active contour lines are presented and an overview of different tracking methods is developed. Based on the literature review the general strategies of different techniques are explored and their possible research directions are analyzed. Literature [19] proposes a Combined Gaussian Hidden Markov Model CGHMM-KF based scheme for monitoring and tracking of targets in a secure area in multi-camera sensor networks. It was experimentally concluded that the method is able to achieve good detection and tracking under a wide range of environmental conditions. Literature [20] discusses the target tracking algorithms in different environments. Color based target tracking detects and counts the motion of single and multiple targets in multiple frames. And further developed a single target tracking algorithm considering the shape and color of the object. Literature [21] aims to provide quality image data for target detection and tracking through noise reduction and image enhancement methods. Noise in the original video is removed by search and match method and block matching technique is used to eliminate undesirable motion in the video. Simulation results specify that these algorithms have good performance in handling noise etc. and can be used to detect and track moving targets in video sequences. Literature [22] initiated the study of key issues in video target detection and recognition, including accurate segmentation of target and background, shadows in complex scenes, etc., and proposed solutions to these problems. Literature [23] developed intelligent transportation video tracking technology by adopting open source hardware video server and software open CV image processing library for the problems exposed in urban transportation system, which realized intelligent traffic monitoring, target trajectory monitoring, intelligent traffic management, etc., and maintained the traffic order. Literature [24] describes the current state of research on tracking methods and categorizes them to identify useful tracking methods. Feature descriptors used in tracking to describe the appearance of the tracked object and object detection techniques are examined. The advantages and disadvantages of the tracking methods were revealed by grouping them. Literature [25] explores the development of algorithms and introduces the definition of video target tracking algorithms, describing the historical context in which they were created. By comparing and analyzing these reveals its performance, characteristics, advantages and disadvantages. The technological advancements and their challenges are also explored in depth, as well as the methods of evaluation of the algorithms and the current state of research, and their future trends are illustrated.

In this paper, on the basis of multi-scale convolutional neural network, combining the video image signal processing technology based on background subtraction method and the principle of multi-target tracking algorithm, the target tracking optimization technology in video image signal processing is proposed. It utilizes the convolutional layer of VGG16 to design the network framework, and all the feature information in the network is extracted, and the extraction method utilizes the residual splicing operation. The dual attention mechanism module is embedded in the feature extraction in order to give more attention to the target location. A spatial attention mechanism module is used to

add more information about the target location to increase the data association accuracy. The feature fusion method also adopts a new approach by using ROI align to integrate the features and finally the improved Hungarian algorithm is used to perform data association operations on the features to get the final tracking results of the algorithm. In the video image signal processing target tracking practice of this paper's target tracking optimization technology to carry out tracking performance experiments, with the OTB database of eight groups of mobile network video as a sample to carry out this paper's target tracking optimization technology application practice.

2. Target tracking optimization technique based on video image signal processing

Target tracking is an important branch of computer vision that utilizes contextual information of a video or image sequence to model the appearance and motion information of a target in order to estimate its location. However, in video image signal processing, unknown targets and tracking scenarios, limited target initial states, and time-varying target appearance make the target tracking task very challenging. In this paper, we focus on the key techniques of target tracking based on convolutional neural networks, aiming to improve the performance of target tracking techniques in video image signal processing.

2.1. Video image signal processing technology

Background subtraction is one of the most commonly used video image signal processing techniques [26]. The idea of background subtraction is to divide the surveillance scene into background and motion foreground, subtract the background from the video image, and the remaining is the motion target. The specific operation is to save the background image first, and then when data acquisition is performed, the current captured video image and the background image are subtracted, and the difference result is binarized to get the region with the value of 1, which is the motion target region.

Establish a mathematical model for, set the current frame image for $f_c(x,y,t)$, t moment background image for $f_b(x,y,t)$, then the current frame image and the background image for the difference and binarization can be expressed as:

$$f_d(x,y,t)=\begin{cases}1 & if \ \left|f_c(x,y,t)-f_b(x,y,t)\right|>T \\ 0 & otherwise\end{cases} \tag{1}$$

Several commonly used background modeling and updating methods are statistical averaging, median filter, Gaussian model modeling, nonparametric modeling, and W4 methods. Each of them is briefly described below:

1) Statistical averaging method. Statistical averaging method for the current continuous acquisition of N frame of the image to accumulate the average to obtain the background image, that is, a continuous sequence of images for statistical averaging, the results of statistical averaging as the background. The mathematical model is as follows: let the knd frame image be f_k and the background before the k th frame image be B_k, then there are:

$$B_k=\frac{1}{N}\left(f_k+f_{k-1}+\cdots+f_{k-N+1}\right) \tag{2}$$

Equation (2) can also be expressed as:

$$B_k=B_{k-1}+\frac{1}{N}\left(f_k-f_{k-N}\right) \tag{3}$$

In Eq. (3), B_{k-1} represents the background image at frame $k-1$, f_{k-N} represents the image at frame $k-N$, and the value of N is generally determined according to the speed of the moving target and the sensitivity of the system.

2) Median filter method [27]. The median filter method utilizes the idea of median filtering to take the middle value of the video image sequence in time as the background. Usually, a moving target does not stay for a long time in the surveillance scene, and the pixel values of each point in the video only change significantly when the target appears. A median filter is set on the pixel values of each point over a period of time, and the middle value is selected as the background of the pixel. The specific implementation process is to cache the most recent N frame of video image, and then select the median value of the pixel at the same position in each frame in the cache as the background value of that pixel. The size of N is generally determined by the motion speed of the target and the shooting sample rate of the camera. The mathematical model is represented as shown in equation (4):

$$B_k(x,y)=median\left(I_{k-1}(x,y),I_{k-2}(x,y),\cdots,I_{k-N-1}(x,y)\right) \tag{4}$$

In Eq. (4), $B_k(x,y)$ represents the background pixel value at pixel point (x,y) in frame k, and $I_k(x,y)$ represents the pixel value at point (x,y) in frame k.

This algorithm needs to store N times the data of the frame image, which is suitable for the surveillance scene, the motion target is not big and the motion is continuous, if the motion target is big and the motion is very slow, the detected result will be empty phenomenon, in order to solve this problem, it is necessary to increase the sliding window, which will need to store more video frames images.

3) Gaussian-based background modeling. In an ideal stationary background, the pixel values of the same pixel point of the captured video image are generally fixed, but due to the variation of environmental factors such as lighting changes, the captured sample values obey a single Gaussian distribution in a statistical sense. The background point is mainly distributed in the center position with high probability in the single Gaussian model, while the motion target is distributed in the region far away from the center of the Gaussian model, which occurs with less probability. Therefore, the average pixel value μ_0 and variance σ_0^2 of the pixel point can be calculated, and the pixel mean of each pixel point is selected as the background B_0. In this way, each pixel point in the background satisfies the Gaussian distribution. The mathematical representation is shown in equations (5), (6), and (7):

$$B_0(x,y)\sim N\left(\mu_0,\sigma_0^2\right) \tag{5}$$

$$\mu_0(x,y)=\frac{1}{M}\sum_{k=0}^{M-1}f_k(x,y) \tag{6}$$

$$\sigma_0^2=\frac{1}{M}\sum_{k=0}^{M-1}\left[f_k(x,y)-\mu_0(x,y)\right]^2 \tag{7}$$

This modeling method is a single Gaussian background modeling method, which is suitable for simple background environments, such as indoor or occasions where there is no random movement (slight shaking) in the scene. However, this modeling approach has limitations if the monitoring scene has multiple forms, such as the swaying of tree branches in the wind. In such occasions, the background can be considered to have several situations, and each Gaussian model corresponds to one of the possible situations in the background, and multiple Gaussian models are utilized for hybrid modeling. The specific modeling method is to assume that there are a total of k possible Gaussian distribution for each pixel point value. Noted as $\eta\left(x,u,\mu_{i,t},\sigma_{i,t}\right)$, $i=1,2,\cdots,k$, respectively, there are:

$$f\left(I_t=u\right)=\sum_{i=1}^{M}\omega_{i,t}\eta\left(x,u,\mu_{i,t},\sigma_{i,t}\right) \tag{8}$$

In Eq. (8), $\eta\left(x,u,\mu_{i,t},\sigma_{i,t}\right)$ is the i rd Gaussian model at moment t, $\mu_{i,t}$ is its mean, $\sigma_{i,t}$ is its standard deviation, and $\omega_{i,t}$ is the weight of the i th Gaussian model. For a pixel in the newly

captured video image, it is first judged which Gaussian distribution it is closest to, and then different Gaussian distribution parameters are updated to update the background model. The update method is as follows. Set the deviation threshold of each model to k, update the parameters of the closest Gaussian model, and update the parameters of the Gaussian model that meets $\left|I_t-\mu_{i,t-1}\right|\le k\sigma_{i,t-1}$, as shown in equations (9), (10), and (11):

$$\omega_{i,t}=(1-\alpha)\omega_{i,t-1}+\alpha \tag{9}$$

$$\mu_{i,t}=(1-\rho)\mu_{i,t-1}+\rho I_t \tag{10}$$

$$\sigma_{i,t}^2=(1-\rho)\sigma_{i,t-1}^2+\rho\left(I_t-\mu_{i,t}\right)^2 \tag{11}$$

If no corresponding Gaussian model is found for the currently captured pixel values, the Gaussian model with the minimum weights is updated, and the other Gaussian models remain unchanged, only their weights are changed, and the values are taken as shown in equation (12). The mean value of the model to be updated is taken as I_t, the deviation is taken as the initialized maximum deviation, and the weights are taken as the initialized minimum weights:

$$\omega_{i,t}=(1-\alpha)\omega_{i,t-1} \tag{12}$$

In the hybrid Gaussian background modeling process, the background model is multimodal as it is a mixture of several single Gaussian models. Such a fuzzy matching strategy is used for the detection of moving targets in the scene.

4) Non-parametric modeling method. The nonparametric modeling method uses a Gaussian kernel nonparameterization to estimate the probability density of the pixel value $I_1,I_2,I_3,\cdots,I_N$ of the most recent N frame of the image. If the estimated value $f(I_t)$ of a pixel value I_t is less than a pre-given threshold, the point pixel is considered to be a motion target pixel. The Gaussian kernel nonparametric estimation model is shown in equation (13):

$$f\left(I_t=u\right)=\frac{1}{N}\sum_{i=t-N}^{t-1}K\left(u-I_t\right) \tag{13}$$

The nonparametric modeling method is able to deal with lighting changes in the background, slight camera jitter, etc., and can quickly adapt to the scene edging. The algorithm requires the same cache as the median filter method in the implementation process, both need to cache N frame of video image, the complexity is about the same as the median filter, but each pixel should be kernel density estimation, the computation consumes more time than the median filter, and its characteristics are not suitable for implementation on hardware FPGA.

5) W4 method. The basic idea of the W4 method is to consider the pixels in the region $\left|I_t(x,y)-I_t^{\max}(x,y)\right|>D_t^{\max}(x,y)$ or $\left|I_t(x,y)-I_t^{\min}(x,y)\right|>D_t^{\max}(x,y)$ as motion target pixels by observing the maximum value $I_t^{\max}(x,y)$, minimum value $I_t^{\min}(x,y)$, and maximum value of the difference between neighboring frames $D_t^{\max}(x,y)$ of each pixel in the video image over a period of time, and using these three quantities to represent the background.

2.2. Principle of multi-target tracking algorithm

Despite the variety of existing multi-target tracking algorithms, the vast majority include some or all of the following processing steps.

1) A detection phase in which the input frames are processed and analyzed using a target detection algorithm to locate and identify the target object.

2) A feature extraction or motion prediction phase, which extracts appearance features or other interaction features of the target detection frame generated by the detector, analyzes the target trajectory and predicts the position of the tracked target in the next frame using a motion predictor.

3) Matching phase, which uses the obtained appearance and motion features to calculate the degree of similarity between the detection frame and the tracked target.
4) Association phase, where the degree of similarity is used to associate the detection frame with the tracked target and assign the same ID to the detection frame as to the tracked target.

Target detection algorithms often choose the $P-R$ curve as an evaluation metric, and the $P-R$ curve refers to a curve made with $(Recall)$ as the vertical coordinate and $(Precision)$ as the horizontal coordinate, where the precision and recall are given by:

$$precision = \frac{TP}{TP + FP} \tag{14}$$

$$recall = \frac{TP}{TP + FN} \tag{15}$$

Where, TP is the number of true positive samples, FP is the number of false positive samples, FN is the number of false negative samples, $recall$ denotes the recall rate and $precision$ denotes the precision rate.

Mean miss rate and F_1 value:

$$Avgmisrate = \frac{FN}{FN + TP} \tag{16}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{(precision + recall)} \tag{17}$$

2.3. Principles of Convolutional Neural Networks

Convolutional layer is a basic layer of the hidden layer in the convolutional neural network, usually contains multiple convolutional kernels, in which each element of the convolutional kernel has its corresponding weight coefficient and bias, the essence of the convolutional operation can be regarded as a number of weighted summation of a number of small regions of the input features, in which the elements of the convolutional kernel are regarded as the weights. The range of the convolution kernel determines the range of these small regions, which is often called the "sensory field". Specifically, during the convolution operation, the convolution kernel slides sequentially over the input features at a certain distance, multiplies the elements of the matrix by each small region in the receptive field, adds them up, and then sums them with the bias. Where the distance over which the convolution kernel slides is referred to as the step size. Equation (18) is the formula for the convolution operation:

$$\begin{aligned} Z^{l+1}(i,j) &= \left[Z^l \otimes w^{l+1} \right](i,j) \\ &+ b(i,j) \in \{0,1,\ldots L_{l+1}\}, L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1 \end{aligned} \tag{18}$$

Where Z^l and Z^{l+1} denote the input and output features of the layer $l+1$ convolution respectively, b is the deviation, $\otimes$ denotes the cross-correlation operation, and L_{l+1} is the size of the feature map Z^{l+1}. $Z(i,j)$ denotes the element at point (i,j) in the feature map, f denotes the size of the convolution kernel, s_0 denotes the sliding step of the convolution kernel, and p denotes the padding parameter of the convolution operation, where the size of the padding parameter represents the number of turns of the 0 value that are padded around the input features during the convolution operation.

2.4. Multi-scale convolutional neural network based target tracking technique

Different features can be extracted from different channels of the convolutional neural network, including shallow texture, color, etc., deeper contour information, and the deepest semantic information, which can be fused to obtain different feature matrices for different targets. These feature

matrices can be the basis for data association afterwards, thus realizing the multi-target tracking algorithm.

Based on the above experimental results, this chapter uses the basic framework of the VGG16 network to design the network structure, and the network structure we designed makes full use of the characteristics of the convolutional neural network that the features extracted from different scales and different channels are different to design the feature extraction method for the target, and then uses the obtained feature matrices as the basis of the data correlation and other operations, and ultimately realizes the multi-target tracking algorithm [28].

2.4.1. Multi-channel and multi-scale feature extraction and fusion. In the whole framework, in order to better extract and fuse the features of multiple channels, two more important modules are added to the model, which are are residual splicing convolution module and ROIalign feature integration module. Meanwhile, in order to achieve higher accuracy, the final correlation operation also uses the improved Hungarian algorithm, and this subsection focuses on analyzing the above three modules [29].

1) Residual splicing convolution module. In the process of deeper convolution of the feature maps of VGG full and extended convolutional networks, it is found that the phenomenon of gradient dispersion occurs during training, and after further experiments, it is found that the convolution layer in the algorithm for convolving the feature maps is too deep, which leads to the disappearance of the gradient. Therefore, this subsection improves the network based on the residual neural network algorithm.

Residual connection means that a feature map after convolution adds the original feature map with the convolved feature map, so that more original details can be retained without loss, thus achieving the purpose of network deepening. Its formula is:

$$y = F\left(x_n\right) + x_n \tag{19}$$

y denotes the output feature map, F denotes the convolution operation, and x_n denotes the layer n feature map. By doing so the number of layers of the convolutional neural network can be effectively deepened.

2) Feature integration module. The multi-channel and multi-scale feature extraction of the algorithm in this chapter refers to the extraction of target features from the eight different scales of the convolutional layers in the network, if the feature information is directly integrated, it will require a huge amount of computation, these feature information is very important for each target, but in order to improve the efficiency of computation, it is necessary to make cuts, so we need to reduce the amount of computation and improve the efficiency of computation.

Using ROIalign, the target position remains unchanged after convolution of the original image, and the relative position of the target box is calculated using the method of bilinear interpolation. The formula is as follows, assuming that we want to know the value of function f at point $P = (x, y)$, we know the value of function f at four points $Q_{11} = \left(x_1, y_1\right)$, $Q_{12} = \left(x_1, y_2\right)$, $Q_{21} = \left(x_2, y_1\right)$ and $Q_{22} = \left(x_2, y_2\right)$, then use the bilinear interpolation can be calculated at the value of $P = (x, y)$ position. The method is as follows.

First calculate the linear interpolation in the x direction to get:

$$f\left(R_1\right) \approx \frac{x_2 - x}{x_2 - x_1} f\left(Q_{11}\right) + \frac{x - x_1}{x_2 - x_1} f\left(Q_{21}\right) where \ \ R_1 = \left(x, y_1\right) \tag{20}$$

$$f\left(R_2\right) \approx \frac{x_2 - x}{x_2 - x_1} f\left(Q_{12}\right) + \frac{x - x_1}{x_2 - x_1} f\left(Q_{22}\right) where \ \ R_2 = \left(x, y_2\right) \tag{21}$$

The linear interpolation in the y direction is:

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} f\left(R_1\right) + \frac{y - y_1}{y_2 - y_1} f\left(R_2\right) \tag{22}$$

Similarly, the final requirement is $f(x, y)$, which is similar to the above.

3) Data association module. The commonly used data association is the Hungarian algorithm i.e. the algorithm for finding the maximum match in graph theory. Here we use the improved Hungarian algorithm to match and correlate the feature information, the specific algorithm is explained as follows. Using the feature information of the target in the current frame and the feature information of the previous n frame (here n will be mentioned in the experimental part, generally take $1 < n < 30$) to do the correlation calculation. In this chapter, Euclidean distance is used to calculate the Euclidean distance between the feature information of each target in the current frame and the feature information of all targets in the previous n frames, and the smallest value obtained is the tracking result. However, if there is a possibility that two targets have similar features resulting in smaller distances, then the data association module, i.e., Hungarian Data Association Module, is needed. In this case, the data association module, i.e., the Hungarian algorithm, is needed.

This chapter adds a loop and judgment before the implementation of the Hungarian algorithm, the loop is to find the closest feature information to the original target among the targets with similar feature information. The judgment is to delete the newly appeared target in the current frame and make it a new tracking trajectory without performing the data association operation with the previous n frames. By adding the above two a priori knowledge to the Hungarian algorithm, the final tracking accuracy of the algorithm in this chapter can be effectively improved and the robustness of the algorithm is also enhanced.

2.4.2. Dual attention mechanisms. This paper introduces a spatio-temporal dual attention mechanism, and the two attention mechanisms are described in detail below.

1) Spatial attention mechanism module. This chapter adds spatial attention module to the network framework to make the network more sensitive to the details of features.

Next, the three resultant maps are spliced based on the channel dimension to obtain a feature map of $H \times W \times 3$, then the spliced resultant map is convolved and the channel dimension is reduced to 1. Finally, the feature map of $H \times W \times 1$ is activated by the sigmoid activation layer to generate a spatial attention feature map, which is then multiplied with the input feature map to obtain the final feature map, which is the enhanced feature map of our target feature. This feature map is the feature map after we enhance the target features. The above operation can also be expressed by the following formula:

$$M_s = \sigma\left(CONV\left(Avg\left(M_i\right) + Max\left(M_i\right) + Rand\left(M_i\right)\right)\right) \times M_i \tag{23}$$

In Eq. M_s denotes the final output spatial attention feature map, M_i denotes the input feature map of the previous layer, Avg, Max and $Rand$ denote the global average pooling, global maximum pooling and global random pooling, respectively, $CONV$ denotes the convolution operation, and σ denotes the $Sigmoid$ -activation layer, and the plus sign in Eq. 1 denotes the feature map stitching operation, and the multiplication sign denotes the matrix multiplication of the feature map (multiplying by columns).

2) Temporal Attention Mechanism Module. This section describes the temporal attention mechanism module. The temporal attention module differs from the spatial attention module in that its implementation is not based on a network structure, but utilizes a programming-like approach for its implementation. Temporal attention mechanism, as the name suggests, is a time-related mechanism, the main principle is to have different levels of attention to different timelines of the sequence in the time series, which is generally mostly used in RNN.Since our tracking object also has a time series (video), we can also implement the temporal attention mechanism in the network.

According to the normal logic, the weights close to the current frame number should be as large as

possible, and those far away from the current frame number should be made small. In accordance with this idea, this method has been experimented to obtain an effective method that can make the weights attenuate, the formula of which is shown below:

$$S_f = \frac{\sum_{i=0}^{n} S_p^{(i)}}{n} + \frac{\sum_{j=0}^{n} S_p^{(j)}}{\sum_{j=0}^{n} j} \tag{24}$$

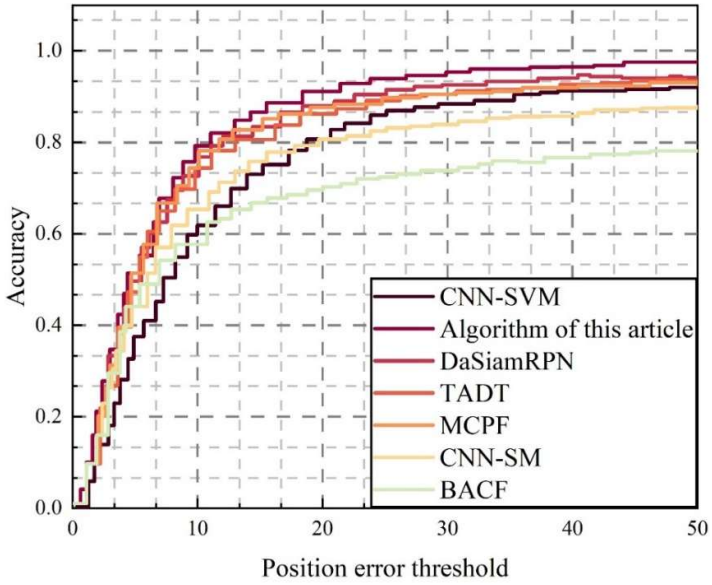
In the above formula (24), S_f indicates the weight size of the target in the current frame, and S_p indicates the weight size of the target in the past frame, according to the formula, the weight of the target in the current frame is able to be calculated, and it can be seen from the formula that its weight value is decayed according to the time, and the closer the target is to the current frame, the higher the weight is, and the further the target is from the current frame, the lower the weight is.

3. Video image signal processing target tracking practices

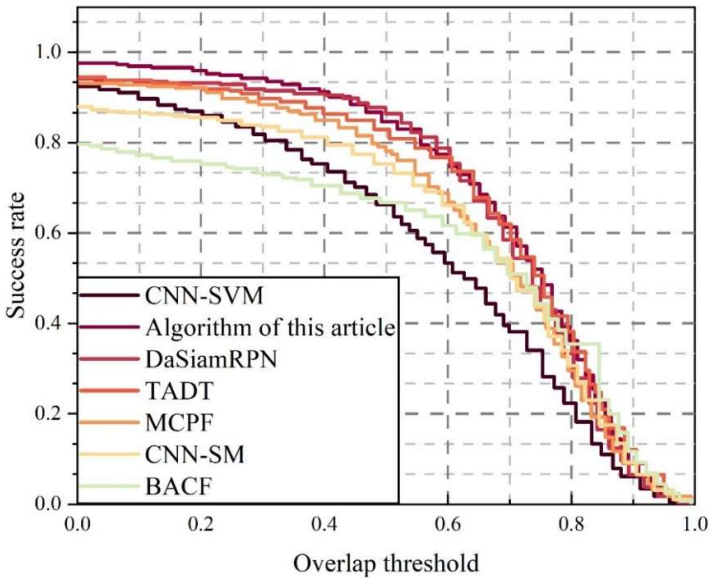
This chapter for the video image signal processing in the target tracking experiment is mainly divided into two parts, one is the proposed target tracking optimization technology performance of this paper for experimental evaluation, the second is the use of video image samples to carry out this paper's target tracking optimization technology application practice.

3.1. Tracking performance experiment

In order to demonstrate the performance of the target tracking optimization technique proposed in this paper, six mainstream algorithms are selected and compared on the OTB2015 dataset. The comparison algorithms are DaSiamRPN, TADT, MCPF, CNN-SVM, CNN-SM, and BACF. The tracking results of each different tracking algorithm on the OTB2015 dataset are specifically shown in Fig. 1, and Fig. (a) and Fig. (b) represent the target tracking accuracy and success rate, respectively. It can be seen that the two algorithms, MCPF and BACF, are categorized as correlation filtering algorithms, while this type of algorithms often use handmade semantic features to construct the observation model of the target as a whole, ignoring the interference of the target's background information, and when the target undergoes deformation, the training samples introduce background pixels, which affects the model performance. DaSiamRPN, TADT, CNN-SM, CNN-SVM belong to deep tracking algorithms, i.e., the use of CNN to build a model to realize the features of image features, but the original CNN model contains more spatial and channel feature redundancy in the extracted features, which does not ensure better feature discrimination and differentiation when comparing and classifying features. In contrast, the algorithm in this paper incorporates a dual attention mechanism in the network and employs a spatial attention mechanism module to increase the data association accuracy and be more discriminating in the features extracted, ensuring better tracking results.



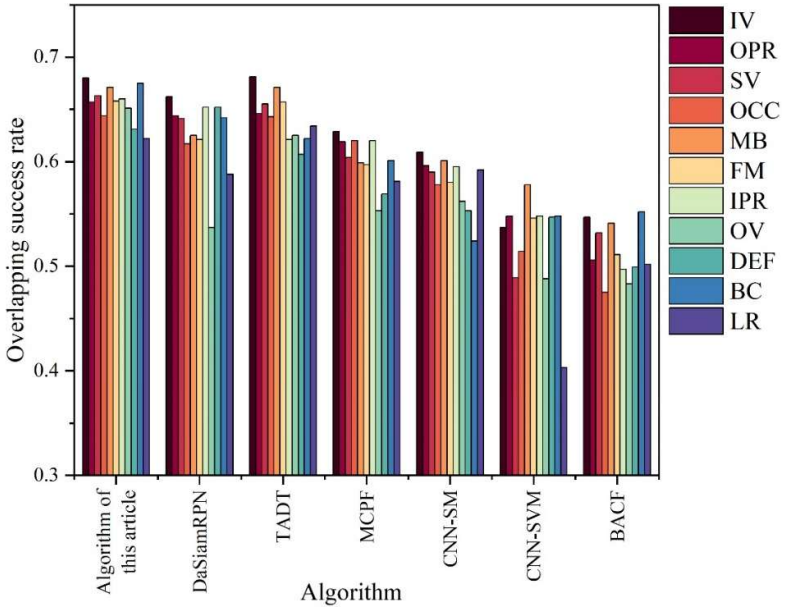
(a)Accuracy



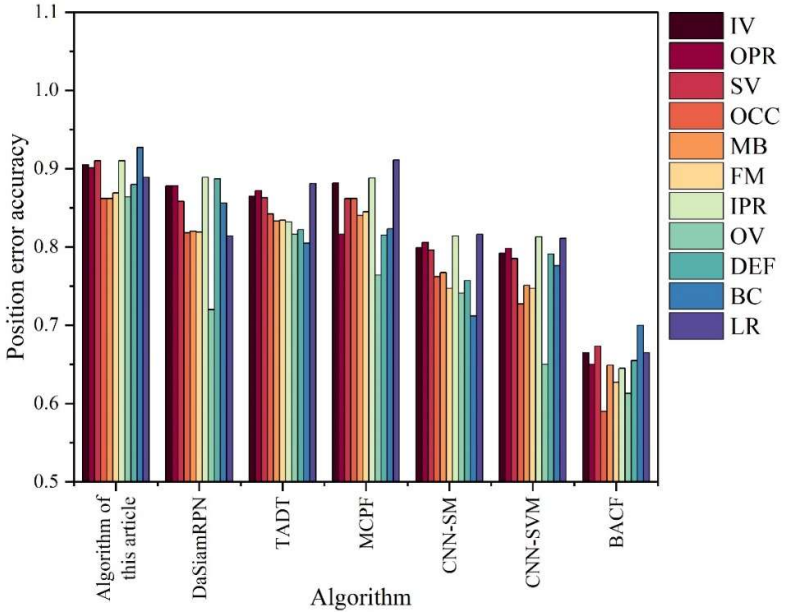
(b)Success rate

Fig. 1. Tracking result

To further evaluate the tracking effectiveness of the proposed algorithms, this chapter also compares 11 tracking scenarios in OTB2015. The tracking results of each comparison algorithm are specifically shown in Fig. 2, with Fig. (a) and Fig. (b) indicating the overlap success rate and position error accuracy, respectively. The algorithms in this chapter achieved better tracking results in several tracking scenarios, especially in the challenging factors such as background clustering (BC), scale variation (SV), and out-of-view (OV). This is due to the fact that the algorithm in this chapter improves the ability of the model to acquire features by focusing on the important information in the target through the dual-attention mechanism module in the absence of the target and background interference, and reduces the number of features for feature fusion by using ROI align in the network, and adopts the improved Hungarian algorithm to get tracking results with the fused features, which ensures tracking results.



(a)Overlapping success rate



(b)Position error accuracy

Fig. 2. Overlapping success rate

3.2. Target Tracking Application Experiment

3.2.1. Analysis of tracking accuracy and success rate. The target tracking optimization technique proposed in this paper is applied to perform target tracking experiments on eight sets of mobile network video images in the OTB database. The tracking application experimental results are compared using this paper's algorithm with DS_v2 and FFT16 target tracking algorithms to compare the tracking accuracy (DP), tracking success rate (SR), and tracking center location error (CLE), respectively. The experimental results of tracking accuracy and tracking success rate of the three methods are shown in Table 1. There is a large gap between the three tracking methods in tracking accuracy, the average tracking accuracy of this paper's method is 97.89%, the average tracking accuracy of DS_v2 method and FFT16 method is 81.5% and 88.22%, this paper's method is higher than

the DS_v2 method and FFT16 method by 16.39% and 9.67%, respectively . The tracking accuracy of all three methods decreases irregularly with the increase of video frames, but the fluctuation of this paper's method is small, which indicates that the method has the stability of mobile network target tracking. In terms of the tracking success rate experimental results, the method in this paper has a higher mobile network target tracking success rate, the average tracking success rate of the method in this paper is 96.02%, the average success rate of the DS_v2 method is 80.53%, and the method in this paper has a higher average success rate than the DS_v2 method by 15.49%.The average success rate of the FFT16 method is 87.42%, and the average success rate of this paper is higher than the average success rate of the FFT16 method. The average success rate of FFT16 method is 8.6% higher.

Table 1. Tracking accuracy and tracking success rate

Video frame number	Tracking accuracy(%)			Target location	Tracking success rate(%)		
	Algorithm of this article	FFT 16	DS_ v2		Algorithm of this article	FFT 16	DS_ v2
100	99.12	89.38	86.62	10	97.22	90.78	82.74
200	96.51	89.24	84.44	20	98.11	87.59	81.6
300	98.69	88.22	82.41	30	95.81	89.36	80.3
400	98.98	92.44	80.81	40	97.16	88.22	81.49
500	98.4	90.55	80.81	50	94.82	89.41	79.76
600	98.83	85.61	78.34	60	96.59	87.67	79.78
700	97.23	86.04	79.79	70	95.01	84.33	78.78
800	95.34	84.31	78.77	80	93.42	82.01	79.82

3.2.2. Analysis of target tracking results. The experimental results of target tracking are shown in Fig. 3. As can be seen from the figure, the tracking results of this paper's method are closer to the actual motion trajectory of the target of the mobile network, and it is known that the average errors of the center position of this paper's method, the DS_v2 method and the FFT16 method are 4.12 mm, 17.84 mm, and 13.26 mm, respectively.Obviously, this paper's method achieves smaller tracking errors, and has a better performance in the target tracking results.

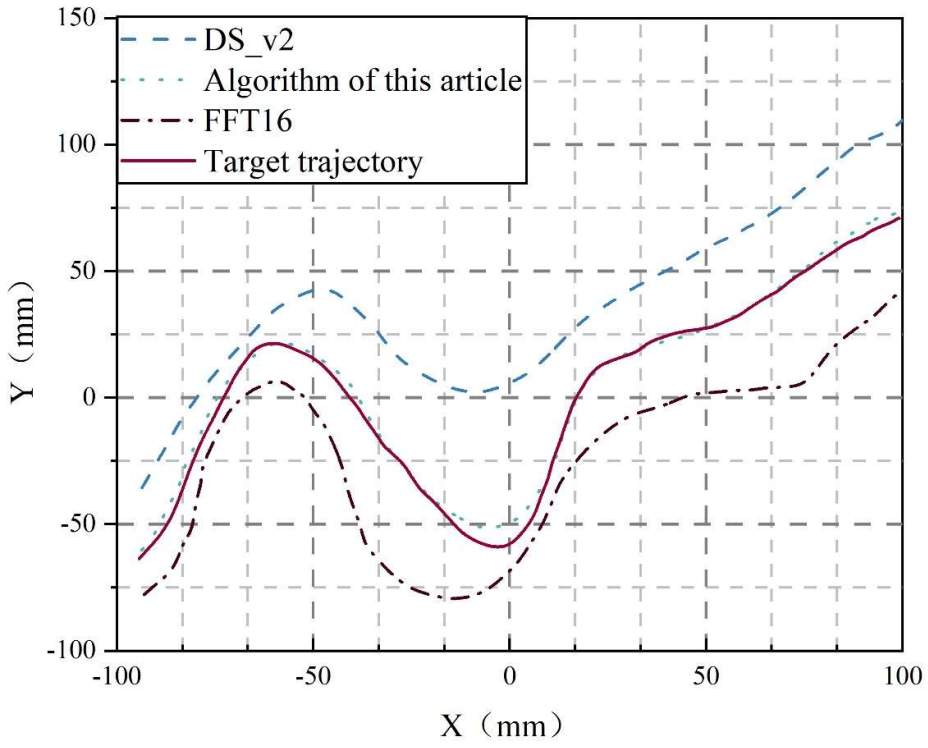
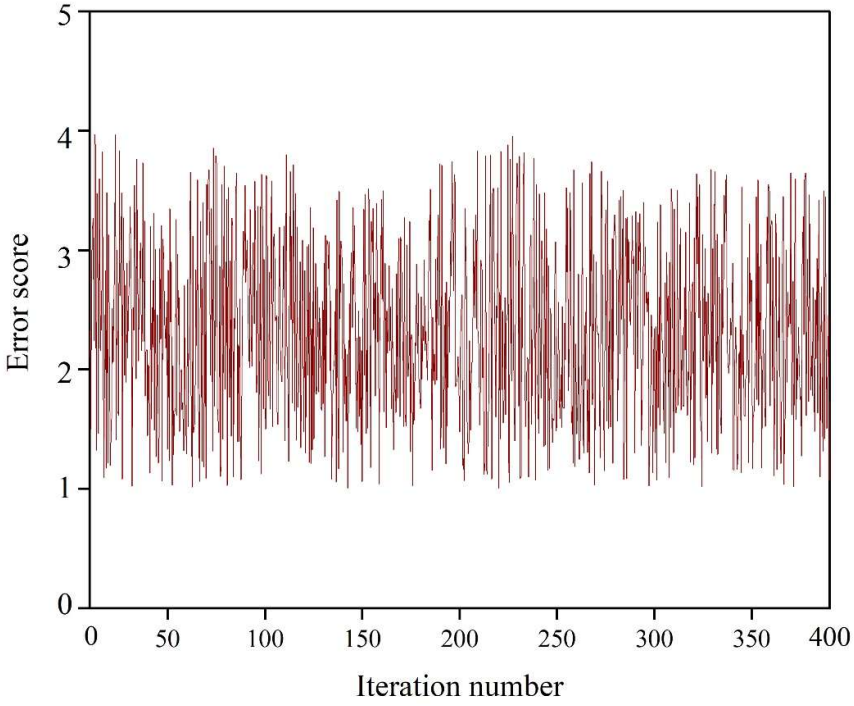
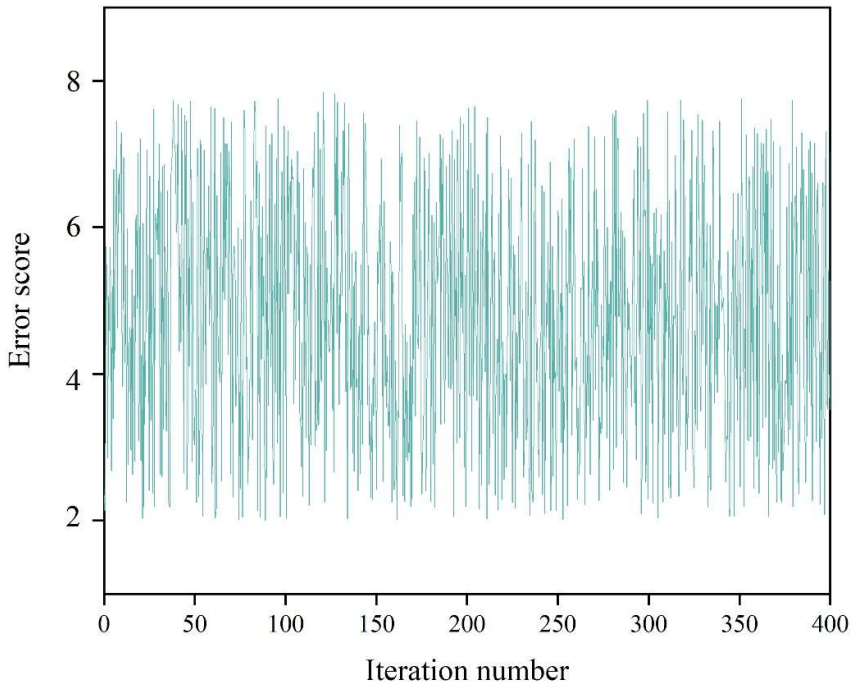


Fig. 3. Target tracking results

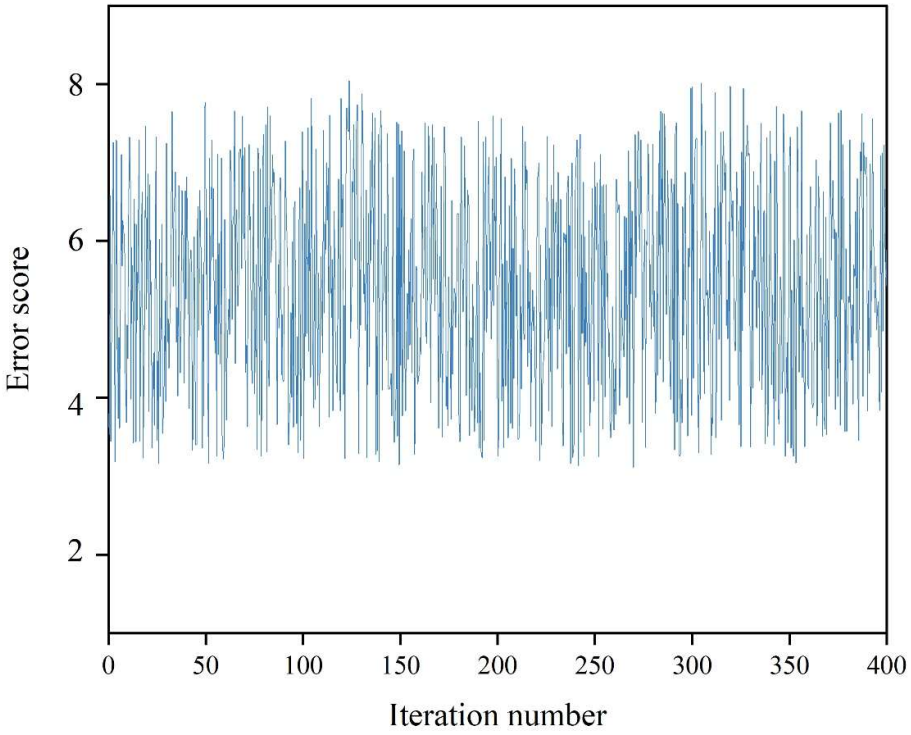
3.2.3. Target tracking error analysis. In this section, the error evaluation metric Root Mean Square Error (RMSE) will be used to assess the gap between the values obtained by the tracking algorithms and the actual position values. The experimental results of the errors of the target tracking positions of the three methods are shown in Fig. 4, Fig. (a), Fig. (b), and Fig. (c) in this way represent the RMSE of this paper's method, DS_v2 and FFT16 methods. The error magnitude of DS_v2 method is the largest, the error magnitude of FFT16 method is the second largest, and the error magnitude of this paper's algorithm is the smallest, which indicates that this paper's method has a better tracking effect in the process of tracking and controlling the target in the video image signal.



(a)Algorithm of this article



(b)DS_v2



(c)FFT16

Fig. 4. RMSE

4. Conclusion

In this paper, based on multi-scale convolutional neural network and with video image signal processing as the research field, we propose a target tracking optimization technique that combines multi-channel and multi-scale feature extraction with dual-attention mechanism. Tracking performance experiments are conducted on the target tracking optimization technique in this paper, comparing with DaSiamRPN, TADT, MCPF, CNN-SVM, CNN-SM, BACF algorithms, this paper's algorithm performs better in terms of tracking results in terms of tracking accuracy and success, and has higher tracking results in terms of challenging factors such as background clustering (BC), scale variations (SV), and out-of-visual-field (OV), and the data correlations are more accurate and have better tracking performance. And in the application practice of target tracking for 8 groups of mobile network video images in OTB database, the average tracking accuracy of this paper's method is higher than DS_v2 method and FFT16 method by 16.39% and 9.67%, and the average tracking accuracy reaches 97.89%. The average tracking success rate is higher than the DS_v2 method and FFT16 method by 15.49% and 8.6%, reaching 96.02%. The tracking results of this paper's method are closer to the actual motion trajectory of the target, and the average error is only 4.12mm, while the average errors of DS_v2 method and FFT16 method reach 17.84mm and 13.26mm. The root mean square error is used to evaluate the difference between the position value obtained by each method and the actual position value, and the method in this paper exhibits the smallest error magnitude. Overall, the target tracking optimization technique in video image signal processing proposed in this paper shows good target tracking performance in both performance tests and application experiments.

References

[1] Guan, L. (Ed.). (2017). Multimedia image and video processing. CRC press.

[2] Liu, Y., Wang, P., & Wang, H. (2018, November). Target tracking algorithm based on deep learning and

multi-video monitoring. In 2018 5th International Conference on Systems and Informatics (ICSAI) (pp. 440-444). IEEE.

[3] Huang, S., Jiang, Y., & Jiang, Y. (2020). Design of target detection and tracking system for sports video. IEEE Access.

[4] Quanan, G., & Yunjian, X. (2020, September). Kalman filter algorithm for sports video moving target tracking. In 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI) (pp. 26-30). IEEE.

[5] Zou, S., Chen, H., Feng, H., Xiao, G., Qin, Z., & Cai, W. (2022). Traffic flow video image recognition and analysis based on multi-target tracking algorithm and deep learning. IEEE Transactions on Intelligent Transportation Systems, 24(8), 8762-8775.

[6] Xu, Y., & Liang, Q. (2021, January). Sports Video Target Tracking Algorithm Based on Optimized Particle Swarm Algorithm. In 2021 International Conference on Information Technology and Contemporary Sports (TCS) (pp. 451-454). IEEE.

[7] Singh, G., & Mittal, A. (2014). Various image enhancement techniques-a critical review. International Journal of Innovation and Scientific Research, 10(2), 267-274.

[8] Qi, Y., Yang, Z., Sun, W., Lou, M., Lian, J., Zhao, W., ... & Ma, Y. (2021). A comprehensive overview of image enhancement techniques. Archives of Computational Methods in Engineering, 1-25.

[9] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., & Lin, C. W. (2020). Deep learning on image denoising: An overview. Neural Networks, 131, 251-275.

[10] Fan, L., Zhang, F., Fan, H., & Zhang, C. (2019). Brief review of image denoising techniques. Visual Computing for Industry, Biomedicine, and Art, 2(1), 7.

[11] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7), 3523-3542.

[12] Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., ... & Gao, M. (2023). Techniques and challenges of image segmentation: A review. Electronics, 12(5), 1199.

[13] Pan, Z., Liu, S., & Fu, W. (2017). A review of visual moving target tracking. Multimedia Tools and Applications, 76, 16989-17018.

[14] Ma, W., & Xu, F. (2021). Study on computer vision target tracking algorithm based on sparse representation. Journal of Real-Time Image Processing, 18(2), 407-418.

[15] Hao, J., Zhou, Y., Zhang, G., Lv, Q., & Wu, Q. (2018, October). A review of target tracking algorithm based on UAV. In 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS) (pp. 328-333). IEEE.

[16] Kumar, M., & Mondal, S. (2021). Recent developments on target tracking problems: A review. Ocean Engineering, 236, 109558.

[17] Ramya, K., Kumar, K. P., & Rao, V. S. (2012). A survey on target tracking techniques in wireless sensor networks. International Journal of Computer Science and Engineering Survey, 3(4), 93.

[18] Ojha, S., & Sakhare, S. (2015, January). Image processing techniques for object tracking in video surveillance-A survey. In 2015 International Conference on Pervasive Computing (ICPC) (pp. 1-6). IEEE.

[19] Vasuhi, S., & Vaidehi, V. (2014). Target detection and tracking for video surveillance. WSEAS Transactions on Signal Processing, 10, 179-188.

[20] Mangawati, A., Leesan, M., & Aradhya, H. R. (2018, April). Object Tracking Algorithms for video surveillance applications. In 2018 international conference on communication and signal processing

(ICCSP) (pp. 0667-0671). IEEE.

[21] Balasubramanian, A., Kamate, S., & Yilmazer, N. (2014). Utilization of robust video processing techniques to aid efficient object detection and tracking. Procedia Computer Science, 36, 579-586.

[22] Pan, Q., & Zhang, H. (2020). Key algorithms of video target detection and recognition in intelligent transportation systems. International Journal of Pattern Recognition and Artificial Intelligence, 34(09), 2055016.

[23] Li, Y., Chu, L., Zhang, Y., Guo, C., Fu, Z., & Gao, J. (2019). Intelligent transportation video tracking technology based on computer and image processing technology. Journal of Intelligent & Fuzzy Systems, 37(3), 3347-3356.

[24] Deori, B., & Thounaojam, D. M. (2014). A survey on moving object tracking in video. International Journal on Information Theory (IJIT), 3(3), 31-46.

[25] Zhou, F. (2024, September). A review on video object tracking algorithms. In Fourth International Conference on Signal Image Processing and Communication (ICSIPC 2024) (Vol. 13253, pp. 234-245). SPIE.

[26] Yuan Dai & Long Yang. (2024). Background subtraction for video sequence using deep neural network. Multimedia Tools and Applications(35),82281-82302.

[27] Yiming He,Yanhua Dong & Hongyu Sun. (2024). Research and Application of the Median Filtering Method in Enhancing the Imperceptibility of Perturbations in Adversarial Examples. Electronics(13),2458-2458.

[28] Souha AYADI & Zied LACHIRI. (2024). Visual Emotion Recognition based on transfer learning technique using VGG16. PRZEGLĄD ELEKTROTECHNICZNY(8),

[29] Juan Pablo Vásconez,Elias Schotborgh,Ingrid Nicole Vásconez,Viviana Moya,Andrea Pilco,Oswaldo Menéndez... & Leonardo Guevara. (2024). Smart Delivery Assignment through Machine Learning and the Hungarian Algorithm. Smart Cities(3),1109-.