IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 December 2023, accepted 23 December 2023, date of publication 25 December 2023,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347405

RESEARCH ARTICLE

Target Perception and Behavioral Recognition Algorithms Based on Saliency and Feature Extraction

WEICHUAN NI[1], BINGTIAN ZHANG[2], JINTING ZHANG[1], JIAJUN ZOU[1], ZHIMING XU[1], AND ZEMIN QIU[1]

[1]Guangzhou Xinhua University, Dongguan 523133, China
[2]Guangdong Hotel Management Vocational and Technical College, Dongguan 523960, China

Corresponding author: Zemin Qiu (qiuzemin@xhsysu.edu.cn)

This work was supported in part by the Guangzhou Science and Technology Project under Grant 202002030273 and Grant 202201011731, in part by the Guangzhou Xinhua College Key Discipline Project under Grant 2020XZD02, in part by the Guangdong Key Discipline Scientific Research Capability Improvement Project under Grant 2021ZDJS144 and Grant 2022ZDJS151, in part by the Young Innovative Talents in Guangdong Ordinary Colleges and Universities Project under Grant 2023KQNCX124, and in part by the Faculty Members of Guangzhou Xinhua University Project under Grant 2020KYYB03.

ABSTRACT In response to the problems of target loss and insufficient accuracy in existing target perception and action recognition algorithms, this paper proposes a target perception and action recognition algorithm based on saliency and feature extraction. This algorithm uses saliency detection techniques to obtain salient regions in images or videos to focus attention on the target. At the same time, feature extraction techniques are combined to extract key nodes and inter-frame correlations from the target information. The experimental results of the measurement data show that this algorithm is superior to traditional detection methods in detecting target behaviors. In addition, it has successfully solved the problems of motion misalignment and jumping in pedestrian detection. Although the node localization of the algorithm needs further improvement, it has shown good application prospects in smart cities and intelligent surveillance. Future work will focus on improving the positioning accuracy of key nodes to enhance its adaptability to different environments and scenarios, providing better support for smart city and other applications.

INDEX TERMS Target perception, behavioural recognition, saliency, feature extraction.

I. INTRODUCTION

With the development of the field of human behaviors recognition, from the initial recognition of simple single movements under limited conditions to today's applications in real natural scenes. Behavioural recognition algorithms are receiving more and more attention. Yang [1] propose the aggregation of squeeze-and-excitation (SE) and self-attention (SA) modules with 3D CNN to analyze both short and long-term temporal action behavior efficiently. Wong et al. [2] this paper proposes a more robust methodology of pedestrian tracking and attribute recognition, facilitating the analysis of pedestrian walking behavior. Specific limitations of a current state-of-the-art method are inferred, based on which several improvement strategies are proposed: 1) incorporating high-level pedestrian attributes to enhance pedestrian tracking, 2) a similarity measure integrating multiple cues for identity matching, and 3) a probation mechanism for more robust identity matching. Muhammad et al. [3] propose a bi-directional long short-term memory (BiLSTM) based attention mechanism with a dilated convolutional neural network (DCNN) that selectively focuses on effective features in the input frame to recognize the different human actions in the videos. Liu et al. [4] propose a human-centered attention mechanism that dynamically highlights regions associated with action recognition according to target appearance to selectively recognize the human-object interaction action.

However, with the rapid development of artificial intelligence and computer vision technology, the increasing

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

demands in the fields of intelligent surveillance, automated driving and human-computer interaction have placed higher requirements on target perception and behaviors recognition algorithms. The traditional algorithms mentioned above suffer from the problems of target action loss and insufficient recognition accuracy, which limit the effectiveness of the algorithms in real-world scenarios. To overcome these limitations, we started to focus on target perception and behaviors recognition algorithms based on saliency and feature extraction.

Unlike the above literature review, to make the algorithm effective for target localization and behaviors analysis. The algorithm in this paper starts from the features of saliency and target information in the image or video, and uses saliency detection techniques to extract the salient region of the target and focus attention on the target [5], [6], [7], [8]. And combines it with inter-frame correlation to ensure the coherence and accuracy of the character's movements, to achieve a comprehensive description and recognition of human behaviors. Furthermore, it avoids the problems of losing the target's action and failing to recognize it with sufficient accuracy. The main contributions of our work are summarized below:

1. Significant areas in pictures or videos: Design salient areas of target information. This algorithm uses saliency detection algorithms to extract salient regions in images or videos, thereby drawing attention to the target of interest. This helps to reduce false positives and improve target detection accuracy.

2. Correlation extraction of significant regions: A correlation extraction mechanism has been developed. This mechanism uses feature extraction techniques to extract the correlation between key nodes and frames of target information. It can better understand the behaviors and intention of the target.

3. Human body model state evaluation and optimization mechanism: During the training process, the state of the human body model is evaluated and optimized, and the extracted features are classified and recognized to achieve target behaviors recognition.

Through a comprehensive evaluation of the target recognition capabilities, we can see that the algorithm in this paper performs well both in real-life scenarios and in the video tests of the dataset. It has a higher action recognition effect and accuracy compared to traditional algorithms. We hope to support the improvement and optimization of target perception and action recognition algorithms through in-depth research and experimental validation. We provide more accurate, reliable and efficient solutions for intelligent systems in real-life application scenarios, thus promoting the development of artificial intelligence and computer vision technology.

The structure of this thesis is as follows: Section II provides an overview of the basic significant object detection model and feature extraction algorithms, and reviews previous work. Section III provides a detailed introduction to the target perception and behaviors recognition algorithm based on saliency and feature extraction proposed in this article. Section IV verifies the performance and effectiveness of the algorithm through experiments. Finally, in Section V provides a summary of the entire article and prospects for future research directions. Through this study, we hope to contribute to the further development of human behaviors recognition.

II. RELATED WORK

In this section, we outline the key concepts in significant object detection model and feature extraction, which are crucial for pedestrian action recognition.

In addition, we have explored previous research specifically targeting pedestrian motion recognition algorithms in smart city scenes.

A. SIGNIFICANT OBJECT DETECTION MODEL

1) BLOCK-BASED MODELS WITH INTRINSIC CUES

The algorithm calculates the contrast value of each pixel point in the image by wrapping around the center of each pixel, thereby understanding the brightness differences in different areas of the image.

$$s(x) = \left\| I_\mu - I_{w_{hc}}(x) \right\|^2 \tag{1}$$

2) REGION-BASED MODELS WITH INTRINSIC CUES

The algorithm incorporates prior conditions and combines PCA (Principal Component Analysis) methods to extract subgroups. Finally, regional comparisons are used to highlight saliency.

$$s(r_i) = \sum_{j=1}^{N} w_{ij} D_r(r_i, r_j) \tag{2}$$

By combining these priors, image processing algorithms can better adapt to complex environments. By combining information such as color prior, center prior, low-rank background, and region contrast to highlight saliency, the accuracy and robustness of image processing algorithms can be improved, making them more responsive to various complex scenes and environments [9], [10], [11].

However, there are some problems with existing saliency detection models.

1. Surrounding contrast by pixel center requires calculation for each pixel point, which requires a large amount of computation, especially for high-resolution images, and the computation time will be very long.

2. Adding prior conditions requires selection and setting based on specific application scenarios, but in practical applications, it is difficult to determine the optimal prior conditions and parameter settings.

Therefore, traditional significance detection models have the following limitations

The adaptability to complex scenes is limited. When dealing with complex scenes, for example when there are multiple overlapping targets in the image or the boundary between the target and the background is blurred, these models may not be able to accurately detect the target.

Manual selection and setting of parameters is required: this requires some understanding and experience of images and application scenarios, and may require multiple trials and adjustments to achieve better results.

In our study, we introduced a motion vector prediction model for adjacent macroblocks in terms of motion saliency, and combined directional search to improve saliency estimation.

B. FEATURE EXTRACTION

The common feature extraction is to combine appearance depth features and motion depth features to form a multi-dimensional feature vector. For convenience, the appearance depth feature and motion depth feature are represented by y_1 and y_2, respectively, and the merged feature is y:

$$y=(\omega_1 y_1, \omega_2 y_2) \tag{3}$$

here ω_i is a weighted coefficient, and $(\omega_1)^2 + (\omega_2)^2 = 1$. Determine the weighting coefficient based on intra class consistency and inter class separability.

Usually, there is a significant variance in sample features within the same class, ensuring that samples within the same nearest neighbor are as close as possible. Assuming $y_i = (\omega_1 y_1^i, \omega_2 y_2^i)$

The sample convenience is i, j. The definition of intra class consistency is:

$$\begin{aligned} S_C &= \sum_{i=1}^{N} \sum_{j \in N_R(F_i)} \frac{< y_i, y_j >}{||y_i|| ||y_j||} \\ &= \sum\nolimits_{i=1}^{N} \sum\nolimits_{j \in N_R(F_i)} \frac{\sum_{k=1}^{2} \omega_k^2 y_i^k y_j^k}{\sqrt{\sum_{k=1}^{2} \omega_k^2 (y_i^k)^2} \sqrt{\sum_{k=1}^{2} \omega_k^2 (y_j^k)^2}} \end{aligned} \tag{4}$$

In the formula, $N_R(F_i)$ represents the index set of the sample [12], [13], [14].

However, there are some problems with existing feature extraction.

1. The computational complexity of intra-class consistency: The computation of intra-class consistency requires the calculation of k nearest neighbors for each sample, which has a high computational complexity.

2. This determination method may depend on specific data sets and application scenarios, and requires certain experiments and adjustments to obtain better results.

Therefore, traditional feature extraction has the following limitations

For complex scenarios, the algorithms need to calculate the neighboring samples of each sample, especially when dealing with large amounts of data, which can lead to a huge demand for computing resources and slow down the processing speed.

Scale sensitivity: When pedestrians show significant size changes in video frames, traditional methods may find it difficult to handle such scale changes and accurately detect and recognize pedestrians.

In our research, we adopt a multi-scale analysis method to extract features from salient regions, and combine feature information such as pedestrian key point correlation, inter-frame correlation, and spatial frequency ratio to achieve accurate recognition of human motion behaviors while reducing algorithm computation.

III. RESEARCH METHODS

In traditional pedestrian action recognition algorithms, target locking is a common approach. However, this paper proposes a target perception and behaviors recognition algorithm based on saliency and feature extraction to overcome the limitations of existing algorithms. This approach effectively exploits target relevance information and integrates adaptive search techniques to better understand target behaviors and intent. By employing these strategies, this thesis aims to improve the accuracy and robustness of the recognition algorithm.

A. CLASSIFICATION: SALIENT SUBSETS AND BACKGROUND SUBSETS

Two subsets of hyperpixels were selected: the salient subset of all hyperpixels $\{x_t\}$ and the background subset. We obtained these two subsets from the hyperpixels in each frame using a method that automatically determines the threshold by using the maximum interclass variance between the target and the background, calculated as follows.

The project derives a method that automatically determines the threshold by using the set of hyperpixels as the set of salient classes with the maximum interclass variance of the background. Assuming that the number of grey levels as i is n_i, the total number of pixels in the image is N, the probability of occurrence of each grey level can be obtained as P_i, then there exists $P_i = n_i/N$. In the segmentation algorithm of an image, a threshold value of λ is used according to the grey level of the image to classify it into two classes. i.e. the salient class C_0 and the background class C_b. However, in practical use, not all histograms are bimodal distributions.

To improve the degree of adaptation between blocks, the probability value coefficient reflects the size of the correlation degree of the image target signal, where the larger the probability value coefficient, the closer the target, and vice versa for the background signal.

The probability value coefficient in the current block is

$$\omega(i,j) = \frac{Max(P_{ij}) - Min(P_{ij})}{\frac{1}{m \times n} \sum_{i,j=0} [\gamma(i,j) \times P_{ij}]} \tag{5}$$

where $Max(P_{ij})$, $Min(P_{ij})$ denote the maximum probability and minimum probability of being selected on the current block. In order to prevent some of the probability value coefficients from being too small and tending to zero, or from being too high due to the probability value coefficient and causing the local optimisation phenomenon, this paper restricts the probability value coefficients in each window in the following way.

In each window, the domain of the probability value coefficients is restricted to the interval $[\omega_{\min}(i,j), \omega_{max}]$. We do

this by making the probability value coefficients $\omega_{\min}(i,j) \leq \omega(i,j) \leq \omega_{max}$ for each small window after each iteration according to this restriction, then the probability value coefficients of the window:

$$\gamma(i,j) = \begin{cases} \gamma_{\min} & \omega(i,j) < \omega_{\min}(i,j) \\ \frac{\sum_{i=1,j=1}^{3,3} \gamma(i,j)}{9} & \omega(i,j) \\ \gamma_{max} & \omega(i,j) > \omega_{max} \end{cases} \tag{6}$$

where the salient class portion $f(x,y) \geq \gamma$ and the background portion $f(x,y) < \gamma$.

Thus, the image is effectively segmented into a series of non-overlapping image subsets. Therefore, the ratio of its salient class to the background occurrence is

$$P_b = \sum_{i=0}^{\lambda} P_i \tag{7}$$

$$P_f = \sum_{i=\lambda+1}^{L-1} P_i \tag{8}$$

In order to improve the coding speed of the algorithm, this paper processes the image by using the form of classification. In order to improve the coding speed of the algorithm, this paper adopts the form of classification to process the image. Because the background of the image in the scene is more complex after the salient class grey level is not much different from the surrounding background grey level, this paper introduces the intraclass variance to further improve the segmentation accuracy of the algorithm. The salient class variance is σ_b^2.

$$\sigma_b^2 = \sum_{i=0}^{s} \sum_{j=0}^{t} [(i,j) - \mu]^2 P_{ij}/P_b \tag{9}$$

The background variance is σ_f^2.

$$\sigma_f^2 = \sum_{i=s+1}^{L-1} \sum_{j=t+1}^{L-1} [(i,j) - \mu_f]^2 P_{ij}/P_f \tag{10}$$

The total image variance is σ_T^2.

$$\sigma_T^2 = \sum_{i=0}^{L} \sum_{j=0}^{L} [(i,j) - \mu]^2 P_{ij} \tag{11}$$

where the grey level of the original image is L.

where is the salient class mean:

$$\mu_b(t) = (\mu_{bi}, \mu_{bj})^T = (\frac{\sum_{i=0}^{s}\sum_{j=0}^{t} iP_{ij}}{P_b(t)}, \frac{\sum_{i=0}^{s}\sum_{j=0}^{t} jP_{ij}}{P_b(t)})^T \tag{12}$$

The background mean:

$$\mu_f(t) = (\mu_{fi}, \mu_{fj})^T = (\frac{\sum_{i=0}^{s}\sum_{j=0}^{t} iP_{ij}}{P_f(t)}, \frac{\sum_{i=0}^{s}\sum_{j=0}^{t} jP_{ij}}{P_f(t)})^T \tag{13}$$

That is, the overall mean of the whole image is μ.

$$\mu = (\mu_i, \mu_j)^T = (\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} iP_{ij}, \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} jP_{ij})^T \tag{14}$$

i.e. the intra-class variance can be obtained as

$$\sigma_\omega^2(T) = P_b\sigma_b^2 + P_f\sigma_f^2 \tag{15}$$

The inter-class variance is $\sigma_B^2(T)$.

$$\sigma_B^2(T) = \sigma_T^2 - \sigma_\omega^2(T) \tag{16}$$

Finally, using normalization, we can obtain:

$$T(s,t) = Arg \max_{0\leq(s,t)\leq L-1} (\frac{\sigma_B^2(T)}{\sigma_B^2(T) + \sigma_\omega^2(T)}) \tag{17}$$

By combining the intra-class variance as a thresholding criterion for image segmentation, we can effectively separate the salient classes.

B. EXTRACTION OF KEY POINTS

In order to eliminate the human body position, body shape and other interference, the algorithm normalizes the human body key points and scale. Human body key points, in this paper, we calculate the set of key points by the 2D points $x_{j,k}$ and $x_{j,k}$ annotated in the image, where $x_{j_1,k}$ and $x_{j_2,k}$ denote the real pixel points corresponding to the two key points j_1 and j_2 for a certain person k in the image.

Calculation: if the pixel point p is close to the annotation points $x_{j,k}$ and $x_{i,k}$, then it reaches the peak of the normal curve, then the pedestrian key node function:

$$S_j^*(p) = \min_k \exp(-\frac{||p - x_{j,k}||_2^2}{\sigma^2}) \tag{18}$$

Set the set of key point positions S_j, the hip centre joint is denoted as S_0, with this joint as the origin, the set of joints S_j is $S_j = S - S_0$ after displacement normalisation.

For the S_j scale normalisation, the processed series J_i is

$$J_i = \frac{\alpha S_j}{\sum ||S_{j_i}||} \tag{19}$$

where α is the scaling factor.

1) OPTIMISING THE MATCHING

Suppose S_j is the detected keypoint location and S_p is the ideal location of the keypoint location.

Then the correlation function between them is R.

$$R = \frac{\sum_{j=1,k=1}^{M,N} S_j(j,k) S_p(j,k)}{\sqrt{\sum_{j=1,k=1}^{M,N} S_j(j,k)^2} \cdot \sqrt{\sum_{j=1,k=1}^{M,N} S_p(j,k)^2}} \tag{20}$$

where M,N is the size of the image, by calculating the mutual correlation function to investigate the matching effect of the image, where the larger the value of the mutual correlation function indicates the better the detection effect.

2) ADAPTIVE MATCHING MECHANISM FOR KEY POINTS

The pedestrian features of the images are extracted from the images in the database and then matched against the pedestrian features of the input test images. If a given set of images has the maximum number of matching points, and this number of points exceeds the set threshold, then

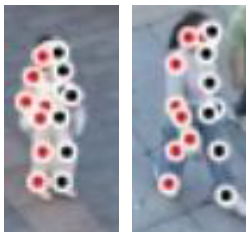

FIGURE 1. Detection charts for key points.

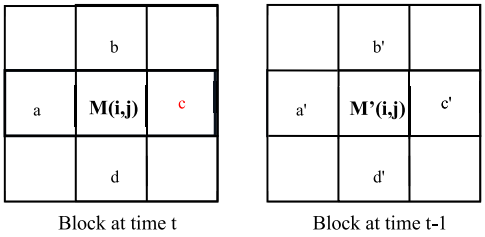

FIGURE 2. Time-dependent motion blocks.

the match is considered successful and information about the corresponding image is returned. By performing a simple segmentation, regions of the human body pose such as feet, hands, head, etc. together determine the final matching result, where each part has its own weight [15], [16], [17].

$$d_w = a * S_{j,a} + b * S_{j,b} + \ldots + n * S_{j,n} \tag{21}$$

where *a,b,c*……and n are the weights of each region.

$S_{j,n}$ is the number of matching points in the nth region for image j in the target 3D library.

Finally, the 3D image matching results and the real image results are combined to determine the final decision factor:

$$d = q * d_1 + (1 - q) * d_2 \tag{22}$$

where d_1 is the 3D image result, d_2 is the real image result, and q is the weight of the 3D image. The adaptivity of the system is mainly reflected in the variation of the weighting factors in the system. As shown in figure 1 detection charts for key points.

C. MOTION TIME CONTINUITY OPTIMIZATION

1) SIGNIFICANCE OF THE CAMPAIGN

In the video sequence, movement objects have strong spatial and temporal correlation, In traditional motion estimation algorithms, mostly spatial correlation in video sequences can be taken into account, by predicting the motion vector of the current block based on the motion vectors of the current block's left neighbor, upper neighbor, upper-left neighbor, and upper-right neighbor, That is, the motion vector prediction model of neighboring macroblocks in the algorithm as shown in figure 2 time-dependent motion blocks.

Let $MV_{i,j}^{t}$ be the motion vector at moment t, and $MV_{i,j}^{t-1}$ is the motion vector at moment t-1 (i.e., the running vector at the previous moment). First calculate the motion-based saliency

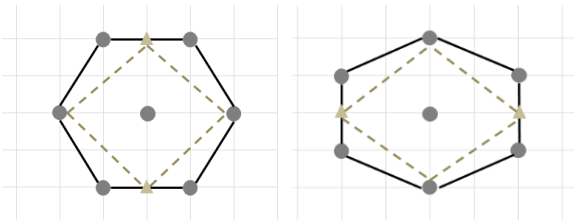

FIGURE 3. Hexagonal search pattern.

of this frame at position.

$$\boldsymbol{S}_t^M(x) = ||\Phi_t(x) - MV_{i,j}^{t}|| \tag{23}$$

where $\boldsymbol{S}_t^M(x)$ is the significance estimated from motion cues. This paper further estimates four additional principal motion vectors to avoid background interference. This is obtained from the mean motion vector stars at the four boundaries (top, bottom, left and right). Then obtain the significance formula for the motion.

$$\boldsymbol{S}_t^M(x) = \min_{\alpha=1:5} ||\Phi_t(x) - (MV_{i,j}^{t})^{\alpha}|| \tag{24}$$

Corresponding to the five main motion vectors (meaning the motion vectors of the whole flow field and the four boundaries). Measure the saliency of pixel x in the frame as its shortest motion distance to these 5 average motion vectors, which in turn greatly improves the significance estimate.

2) DIRECTIONAL SEARCH MODE

Adopts circular search pattern, Circular search modes include triangle search mode, square search mode, and hexagonal search mode. The hexagonal shape has the largest search center interval and the widest coverage, which is conducive to a more accurate search.

The distribution of search points in the data, then the search origin is judged first. if it is not the optimal point, the positive hexagonal search pattern is divided into a horizontal hexagonal search pattern and a vertical search pattern, as shown in figure 3 hexagonal search pattern.

In order to better search for the video, this paper uses a threshold judgment algorithm to select it as follows.

Let the relative bivalent distance between the horizontal and vertical directions be L_x, L_y.

$$L_x = \sqrt{(\sum_{i=0}^{N-1} MV_{ix} - MV_{\text{predx}})/N} \tag{25}$$

$$L_y = \sqrt{(\sum_{i=0}^{N-1} MV_{iy} - MV_{\text{predy}})/N} \tag{26}$$

where N is the number of elements in the set of motion vectors and A and B are the horizontal and vertical components of, A and B are the components of M in the horizontal and vertical directions. By introducing a threshold function, $\lambda_3 = \sqrt{(a^2 + b^2)/2}$, where a, b are the number of long and wide pixels of the current block. When it is in $L_x < \lambda_3 < L_y$, it means that it is relatively intense in the vertical direction, i.e., the vertical hexagonal search pattern can be used to

search in this range. When it is in $L_x < \lambda_3 < L_y$, it means that it is relatively intense in the horizontal direction, i.e., the horizontal hexagonal search pattern can be used for searching in this range.

3) INTERFRAME CORRELATION

The spatial frequency of an image is a metric related to the gradient, it reflects the degree of marginalization of the image, and can well describe the information of the gray scale mutation of the image. Where a greater spatial frequency of the image indicates a more active and clearer image. The spatial frequency of an image for $M \times N$ is defined as $SF = \sqrt{RF^2 + CF^2}$, where RF and CF denote the row and column frequencies, respectively.

$$RF = \sqrt{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=2}^{N}(f(x,y) - f(x,y-1))^2} \quad (27)$$

$$CF = \sqrt{\frac{1}{MN}\sum_{x=2}^{M}\sum_{y=1}^{N}(f(x,y) - f(x-1,y))^2} \quad (28)$$

where the salient class, Part A, is not difficult to derive from the definition above, The spatial frequency centered on (x,y) with M×N as the region size is $SF_i(x,y) = \sqrt{RF_i^2(x,y) + CF_i^2(x,y)}$, where i takes the values of 1 and 2, and $RF_i^2(x, y)$ and $CF_i^2(x, y)$ denote the row frequency domain column frequency of the image i in, respectively.

By performing spatial frequency calculations on the previous frame and the current frame, we can get.

The spatial frequency of the previous frame is $SF_1(x, y)$.

$$SF_1(x,y) = \sqrt{RF_1^2(x,y) + CF_1^2(x,y)} \quad (29)$$

The spatial frequency of the current frame is $SF_2(x, y)$.

$$SF_2(x,y) = \sqrt{RF_2^2(x,y) + CF_2^2(x,y)} \quad (30)$$

In turn, you can define the ratio of the spatial frequency of the previous frame to the current frame:

$$M_{1,2}(x,y) = \frac{SF_1(x,y)}{SF_2(x,y)} \quad (31)$$

The degree of correlation between two frames is very much measured based on the spatial frequency ratio of the resulting image.

In order to test the tracking effect of the algorithm, this paper selects dynamic pedestrians for target localization, verifies that the video with variable scale is examined, and records the detection results in real time as shown in figure 4 effectiveness of target tracking (scale variation).

It can be observed that the algorithm is effective in localizing and tracking a single line of people, and by incorporating inter-frame correlation, the algorithm is able to adapt to scale changes of the target and is robust to continuous attitude changes of the target.

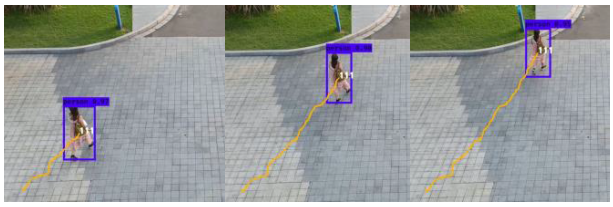

FIGURE 4. Effectiveness of target tracking (scale variation).

D. CONDITION ASSESSMENT AND OPTIMIZATION OF MANNEQUINS

1) EDGE MATCHING ERROR

The edge matching error E_b can be viewed as the Chamfer distance from the edge of the model projection to the edge of the image [18], [19], [20]. Assume that the model projects an edge profile of r(s), $0<s<1$ for a given human body state x_t The corresponding edge profile in the image is z(s), There exists a mapping g(s) which associates each point z(s) on the edge curve in the image with the corresponding point r(g(s)) on the edge curve of the model projection. Thus the edge matching function a natural definition.

$$E_b = \frac{1}{c}\int_0^1 \min(||z_1(s) - r(s)||, u)ds \quad (32)$$

Here c is the gauge normalization constant, which is generally replaced by the length of the edge curve. $z_1(s)$ is the most similar correlation feature to r(s).

$$z_1(s) = z(s') \quad (33)$$

$$s' = arg\min_{s' \in g'(s)} ||r(s) - z_1(s)|| \quad (34)$$

It can eventually be discretized into a summation form.

$$E_b = \frac{1}{cM} min(||z_1(s_m) - r(s_m)||, u) \quad (35)$$

where m is the spatial scaling factor. Here $s_m = m/M$, the above equation can be viewed as finding M points on the edge curve of the model projection, Calculate the distance from each point to the corresponding feature point in the image, E_b is the average of these distances.

2) REGIONAL MATCHING ERROR

We further consider the region matching error [21], [22]. When the model is projected onto the image plane and matched to the image data, the area of the model projection is divided into two parts. A portion of the overlap with the image data is denoted as P1 and the remaining portion is denoted as P2, so the region matching error is E_r.

$$E_r = \frac{|P_2|}{(|P_1| + |P_2|)} \quad (36)$$

Here P_i=(1,2) denotes the area, where it can be replaced by the number of pixel points in the corresponding region. As shown in figure 5 diagram of the areas divided.

FIGURE 5. Diagram of the areas divided.

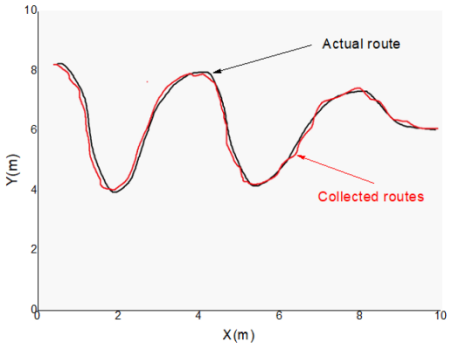

FIGURE 6. Comparison of walking track (top view).

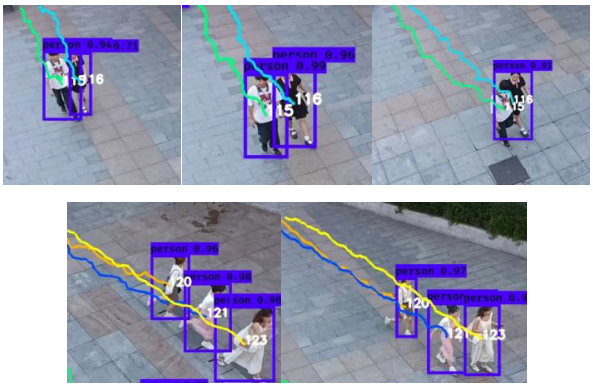

FIGURE 7. Chart of target and trajectory records in real life scenarios.

3) ATTITUDE EVALUATION FUNCTION

The attitude evaluation function can be improved in accuracy and robustness by edge information and region information. Therefore, the algorithm will consider both errors E_b and E_r when applying the attitude function. And use the function $\rho_i(s, \sigma)$.

$$\rho_i(s, \sigma) = ve^{\frac{-s}{\sigma^2}} \tag{37}$$

which in turn accomplishes the conversion of into error to similarity.

$$S(P) = ve^{\frac{-(\alpha * E_b + (1-\alpha) * E_r)}{\sigma^2}} \tag{38}$$

Here P is the state of the mannequin, and the weights of E_b and E_r are adjusted by α.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, the algorithm is implemented on MATLAB, PyCharm platform, to verify the effectiveness of the algorithm. The experimental computer configuration: processor AMD Ryzen 7 5800H (RAM 3.20 GHz), onboard RAM 32.0 GB, graphics card 3060.

To further test the effectiveness of the proposed target detection method. The article selects a real-world scenario and evaluates the algorithm using the algorithm of this paper to perform operations such as motion trajectory extraction and human posture estimation. Meanwhile, in order to test the space and time complexity of the algorithm, the article selects the multi-object tracking dataset, which consists of multi-object scenarios with challenges. This dataset also focuses on crowded scenarios with videos of up to 246 people in a single frame. And compare the literature algorithm [9] and literature algorithm [12] with the algorithm proposed in this paper. The results are shown below.

A. TRAJECTORY DEVIATION COMPARISON

Trajectory deviation comparison is a method of comparing the degree of deviation between different trajectories. In the target localization and trajectory recording experiments, the data of the actual trajectory is collected. The trajectory data of this paper's algorithm is compared and analyzed, and the collected data is shown in figure 6 Comparison of walking track (top view).

By looking at the data of the actual trajectory and the data curve of the algorithmic trajectory in this paper, we can see that there may be some differences between them. However, as long as the deviation is small, the algorithm in this paper can be considered to be quite accurate in predicting the actual trajectories. It is shown that the algorithm in this paper has high accuracy in target localization, which provides strong support for further research and applications.

B. TARGETING AND TRACKING

In order to demonstrate the method discussed in this paper intuitively from the detection performance, this paper selects the actual scene for recording, using this paper's algorithm to set up the identification of the street pedestrians and record their walking trajectories, As shown in figure 7 chart of target and trajectory records in real life scenarios.

By observing the detection effect of the actual scene, we can see that in the actual scene, the algorithm in this paper shows good tracking effect and clear trajectory records, and there will be no problems such as target jumping or misalignment during the recognition process, which indicates that the algorithm has high accuracy and reliability in practical application.

C. COMPARISON EXPERIMENT OF TIME AND SPACE

In this paper, the multi-target scene video MOT20-03 from the Multi-Object Tracking (MOT) dataset is selected. The video has an average of 130.42 pedestrians per frame. The video is 1 minute 36 seconds long and contains 25 frames

Algorithm

Input: video, training model

Parameter: $M, N, i, j, k, P, s, u, v, \alpha, \sigma, \mathrm{S_p}, S_j, E_r = \frac{|P_2|}{(|P_1|+|P_2|)}, E_b = \frac{1}{cM} min(||\mathrm{z_1}\ (\mathrm{s}_m) - \mathrm{r}\ (\mathrm{s}_m)||\ , u).$

Output:

```
def target_perception_and_behavior_recognition(image, frames):
    salient_areas = extract_salient_areas(image)
    correlation_regions = extract_correlation_regions(frames)
    body_model = initialize_body_model()
    R = ∑_{j=1,k=1}^{M,N} S_j(j,k)S_p(j,k) / (√(∑_{j=1,k=1}^{M,N} S_j(j,k)^2) · √(∑_{j=1,k=1}^{M,N} S_p(j,k)^2))
    recognized_behavior = classify_recognized_behavior(
        optimize_body_model(body_model, salient_areas, correlation_regions) )
    S(P) = ve^{−(α∗E_b+(1−α)∗E_r)/σ^2}
return recognized_behavior
```

TABLE 1. Time and space comparison experimental data table.

Algorithm	Processing time per frame (ms)	Total processing time (s)	Memory usage (MB/frame)
Literature Algorithm[9]	20	48	50
Literature Algorithm[12]	25	60	70
Article Algorithm	21	50	40

per second. Different algorithms are run on the same hardware platform and the runtime and memory usage of each algorithm is recorded.

By recording the running time of the algorithms, the time taken by different algorithms to process the same number of targets and behaviors is analyzed. Also, by recording the memory usage of the algorithms, the amount of memory used by different algorithms to process the same number of targets and behaviors is analyzed. The data is shown in table 1 below.

This can be seen from the data in the table above: In terms of running time: this paper's algorithm is comparable to literature algorithm [9] and literature algorithm [12] takes the most time.

In terms of memory usage: this paper's algorithm has the lowest memory usage and literature algorithm [12] has the highest. It can be seen that this paper's algorithm has a lower computational complexity in terms of space.

Considering the time and space complexity together, it can be concluded that the algorithm in this paper is able to maintain low computational complexity while ensuring low space occupancy, which effectively accepts the feasibility of the algorithm.

D. GENERATED BODY POSES

In this paper, we simulate the actual scene and generate human pose data through simulation in order to be used for motion capture and human pose recognition, and then analyze the accuracy of pose recognition and the precision of motion capture. As shown in figure 8 Motion capture rendering (3D).

FIGURE 8. Motion capture rendering (3D).

TABLE 2. Mean values of PCP.

Movement joints	PCP (%)		
	Literature Algorithm[9]	Literature Algorithm[12]	Article Algorithm
Head	85.5	82.2	86.1
Neck	90.4	91.1	91.5
Shoulder	759	79.4	80.5
Arms	81.2	83.1	90.4
Waist	88.5	90.6	91.9

By generating human poses we can see that the simulated human pose generation method achieves good results in terms of accuracy and realism. The algorithms in this paper are able to accurately capture the movements and poses of the characters and generate the corresponding 3D models. The algorithms are all able to effectively reconstruct the character's pose, showing high accuracy and stability. Algorithms show good performance and applicability.

E. TESTING OF PCP

To assess the accuracy and error of the algorithms in pose reconstruction, in this paper, PCP (Percentage of Correct Parts) is used to assess the accuracy of the pose estimation algorithms in detecting and localizing the critical points of the human body. In this paper, the pose data generated by the algorithm is recorded by calculating the head, neck, shoulder, arm, waist, leg, etc. The data is shown in table 1 below.

According to the results in Table 2, it can be seen that the algorithm proposed in this paper has relatively high detection accuracy in multiple body parts, outperforming the literature algorithms. These results indicate that the algorithm proposed

in this paper has potential in the field of multi-objective motion analysis and can serve as an effective method to handle related problems.

V. CONCLUSION

In this study, We propose a target perception and behavioral recognition algorithm based on saliency and feature extraction. The algorithm is based on feature information and image saliency, and the algorithm achieves the perception and recognition of pedestrians in images or videos by reclassifying the information using interclass variance, optimizing the matching mechanism in combination with extracting the correlation of inter-frame information, and finally optimizing the human model pose. The experimental results show that the algorithm has high accuracy and robustness, and can effectively identify the behaviors of targets. However, compared to traditional methods, this algorithm still has room for optimization in terms of time complexity. The future research direction will focus on further investigating this aspect to gain deeper insights.

REFERENCES

[1] J. W. Yang, "Target tracking and recognition of a moving video image based on convolution feature selection," *Int. J. Biometrics*, vol. 13, nos. 2–3, pp. 180–194, 2021.
[2] P. K.-Y. Wong, H. Luo, M. Wang, P. H. Leung, and J. C. P. Cheng, "Recognition of pedestrian trajectories and attributes with computer vision and deep learning techniques," *Adv. Eng. Informat.*, vol. 49, Aug. 2021, Art. no. 101356, doi: 10.1016/j.aei.2021.101356.
[3] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Gener. Comput. Syst.*, vol. 125, pp. 820–830, Dec. 2021, doi: 10.1016/j.future.2021.06.045.
[4] S. Liu, Y. Li, and W. Fu, "Human-centered attention-aware networks for action recognition," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10968–10987, Dec. 2022, doi: 10.1002/int.23029.
[5] J. Li, W. Ji, M. Zhang, Y. Piao, H. Lu, and L. Cheng, "Delving into calibrated depth for accurate RGB-D salient object detection," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 855–876, Apr. 2023.
[6] M. Shokri, A. Harati, and K. Taba, "Salient object detection in video using deep non-local neural networks," *J. Vis. Commun. Image Represent.*, vol. 68, Apr. 2020, Art. no. 102769.
[7] K. Qian, P. Chen, and D. Zhao, "GOMT: Multispectral video tracking based on genetic optimization and multi-features integration," *IET Image Process.*, vol. 17, no. 5, pp. 1578–1589, Apr. 2023, doi: 10.1049/ipr2.12739.
[8] S. Gao, J. Yun, Y. Zhao, and L. Liu, "Gait-D: Skeleton-based gait feature decomposition for gait recognition," *IET Comput. Vis.*, vol. 16, no. 2, pp. 111–125, Mar. 2022.
[9] P. Peng, K.-F. Yang, and Y.-J. Li, "Global-prior-guided fusion network for salient object detection," *Exp. Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116805, doi: 10.1016/j.eswa.2022.116805.
[10] S. Zhao, M. Chen, P. Wang, Y. Cao, P. Zhang, and X. Yang, "RGB-D salient object detection via deep fusion of semantics and details," *Comput. Animation Virtual Worlds*, vol. 31, nos. 4–5, p. 1950, Jul. 2020.
[11] D. Zhang and A. Zakir, "Top–down saliency detection based on deep-learned features," *Int. J. Comput. Intell. Appl.*, vol. 18, no. 2, pp. 1–10, Jun. 2019, doi: 10.1142/S1469026819500093.
[12] K. Yang, J. Li, S. Dai, and X. Li, "Multiscale features integration based multiple in single out network for object detection," *Image Vis. Comput.*, vol. 135, Jul. 2023, Art. no. 104714.
[13] X. Liu and J. Yin, "Stacked residual blocks based encoder–decoder framework for human motion prediction," *Cognit. Comput. Syst.*, vol. 2, no. 4, pp. 242–246, Dec. 2020, doi: 10.1049/ccs.2020.0008.
[14] M. Hao, F. Yuan, J. Li, and Y. Sun, "Facial expression recognition based on regional adaptive correlation," *IET Comput. Vis.*, vol. 17, no. 4, pp. 445–460, Jun. 2023.
[15] J. Zhong, L. Jin, and Q. Mao, "Real-time recognition of human motions using multidimensional features in ultrawideband biological radar," *IET Biometrics*, vol. 11, no. 1, pp. 1–9, Jan. 2022, doi: 10.1049/bme2.12038.
[16] G. Li, F. Liu, Y. Wang, Y. Guo, L. Xiao, and L. Zhu, "A convolutional neural network (CNN) based approach for the recognition and evaluation of classroom teaching behavior," *Sci. Program.*, vol. 2021, pp. 1–8, Oct. 2021, doi: 10.1155/2021/6336773.
[17] Y. Lin, W. Chi, W. Sun, S. Liu, and D. Fan, "Human action recognition algorithm based on improved ResNet and skeletal keypoints in single image," *Math. Problems Eng.*, vol. 2020, pp. 1–12, Jun. 2020, doi: 10.1155/2020/6954174.
[18] C. Wang, B. Wang, H. Liang, J. Zhang, W. Huang, and W. Zhang, "W-trans: A weighted transition matrix learning algorithm for the sensor-based human activity recognition," *IEEE Access*, vol. 8, pp. 72870–72880, 2020, doi: 10.1109/ACCESS.2020.2984456.
[19] C.-Y. Luo, S.-Y. Cheng, H. Xu, and P. Li, "Human behavior recognition model based on improved EfficientNet," *Proc. Comput. Sci.*, vol. 199, pp. 369–376, Jan. 2022, doi: 10.1016/j.procs.2022.01.045.
[20] F. Anvarov, D. H. Kim, and B. C. Song, "Action recognition using deep 3D CNNs with sequential feature aggregation and attention," *Electronics*, vol. 9, no. 1, pp. 147–159, Jan. 2020, doi: 10.3390/electronics9010147.
[21] Y. Zhang, Y. Huang, X. Sun, Y. Zhao, X. Guo, P. Liu, C. Liu, and Y. Zhang, "Static and dynamic human Arm/Hand gesture capturing and recognition via multiinformation fusion of flexible strain sensors," *IEEE Sensors J.*, vol. 20, no. 12, pp. 6450–6459, Jun. 2020, doi: 10.1109/JSEN.2020.2965580.
[22] M. He, G. Song, and Z. Wei, "Human behavior feature representation and recognition based on depth video," *J. Web Eng.*, vol. 19, nos. 5–6, pp. 883–902, 2020, doi: 10.13052/jwe1540-9589.195614.