applied sciences

MDPI

Article

Behavioral Parameter Field for Human Abnormal Behavior Recognition in Low-Resolution Thermal Imaging Video

Baodong Wang [1,2,3], Xiaofeng Jiang [1,2,3], Zihao Dong [1,2,3,*] and Jinping Li [1,2,3,*]

[1] School of Information Science and Engineering, University of Jinan, Jinan 250022, China; wangbd@stu.ujn.edu.cn (B.W.); jiangxf@mail.ujn.edu.cn (X.J.)
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China
[3] Shandong College and University Key Laboratory of Information Processing and Cognitive Computing in 13th Five-Year, Jinan 250022, China
* Correspondence: ise_dongzh@ujn.edu.cn (Z.D.); ise_lijp@ujn.edu.cn (J.L.)

Abstract: In recent years, thermal imaging cameras are widely used in the field of intelligent surveillance because of their special imaging characteristics and better privacy protection properties. However, due to the low resolution and fixed location for current thermal imaging cameras, it is difficult to effectively identify human behavior using a single detection method based on skeletal keypoints. Therefore, a self-update learning method is proposed for fixed thermal imaging camera scenes, called the behavioral parameter field (BPF). This method can express the regularity of human behavior patterns concisely and directly. Firstly, the detection accuracy of small targets under low-resolution video is improved by optimizing the YOLOv4 network to obtain a human detection model under thermal imaging video. Secondly, the BPF model is designed to learn the human normal behavior features at each position. Finally, based on the learned BPF model, we propose to use metric modules, such as cosine similarity and intersection over union matching, to accomplish the classification of human abnormal behaviors. In the experimental stage, the living scene of the indoor elderly living alone is applied as our experimental case, and a variety of detection models are compared to the proposed method for verifying the effectiveness and practicability of the proposed behavioral parameter field in the self-collected thermal imaging dataset for the indoor elderly living alone.

Keywords: abnormal behavior recognition; low-resolution thermal imaging; behavioral parameter field; YOLOv4

Citation: Wang, B.; Jiang, X.; Dong, Z.; Li, J. Behavioral Parameter Field for Human Abnormal Behavior Recognition in Low-Resolution Thermal Imaging Video. *Appl. Sci.* **2022**, *12*, 402. https://doi.org/10.3390/app12010402

Academic Editor: Vera Yuk Ying Chung

Received: 13 December 2021
Accepted: 28 December 2021
Published: 31 December 2021

1. Introduction

Currently, thermal imaging cameras are widely used in many fields [1] and have obvious advantages over visible light in intelligent surveillance. Thermal imaging cameras rely on the heat emitted by the object itself to take pictures, which is more suitable for surveillance scenarios that require privacy protection and temperature measurement. Although thermal and visible light images are vastly different in terms of imaging principles, the research idea of behavior recognition is similar. Figure 1 shows an example of a visible light image and thermal imaging image.

(**a**) (**b**) (**c**)

Figure 1. Examples of imaging effects. (**a**) Visible light image. (**b**) Thermal imaging image taken in the same scene as (**a**). (**c**) Image of a person suddenly falling down.

Traditional video behavior recognition methods are mainly based on visual feature extraction or motion feature extraction. One of the methods based on visual feature extraction is mainly by sampling video frames, manually designing features that can characterize actions for the sampled points, and then encoding and processing them for action classification by a classifier. Such methods mainly include two different types of feature extraction: global feature-based [2,3] and local feature-based [4,5] methods. The method based on motion feature extraction mainly uses optical flow fields to extract motion trajectories and then encodes the features for training to complete the classification. One of the most robust algorithms is the improved dense trajectories (iDT) algorithm [6,7]. With the improvement in computer performance, behavior recognition methods based on deep learning have become popular. There are three main types of mainstream methods: dual-stream networks [8–10], 3D convolutional networks [11–13], and behavior recognition methods based on skeletal keypoints [14–17], which have emerged in recent years. Among them, the behavior recognition method based on skeletal keypoints shows excellent results by expressing the spatial structure of different behaviors with the help of graph convolutional networks. The literature [14] proposed the ST-GCN algorithm based on the OpenPose algorithm. Fang et al. [17] first used the target detection algorithm to obtain the human bounding box and then used the AlphaPose algorithm to obtain the sequence of keypoints, and then classified the predicted behavior.

We found that most current thermal imaging cameras are fixed-position gun cameras rather than rotating dome cameras with gimbals. The video resolution is generally low, which leads to the inability to maintain high accuracy and robustness even with the skeletal keypoints-based approach. In this work, we work on the implementation of a method for human behavior recognition in low-resolution thermal imaging video based on fixed scenes, called the behavioral parameter field (BPF). The contributions of this paper are as follows:

(1) The BPF algorithm based on a low-resolution thermal imaging camera is proposed to learn the behavioral features of the human body using features, such as aspect ratio, height, velocity, acceleration, the center of mass, and orientation angle, and then the stability and validity of the behavioral parameter field are ensured by a similarity update strategy. The feature matching method based on intersection and ratio is proposed to identify abnormal behavior.

(2) Considering the high real-time requirements of the surveillance task, the single-stage YOLOv4 [18] algorithm is selected as the baseline in this paper. The effects of three variants of the scSE attention mechanism [19] embedded into the backbone of the model on the performance of the algorithm are studied and analyzed separately to enhance the information representation capability of the YOLOv4 algorithm for small target feature maps.

(3) Taking an indoor solitary elderly scene as an example, we use the method of this paper to complete the recognition tasks of falling down and long-time immobility behaviors, and make a comparative experiment with the two kinds of attitude estimation algorithms on our data set, which shows the effectiveness and practicability of this method in the real scene of a fixed thermal imaging video.

This paper is organized as follows. In Section 2, we first introduce the work related to the behavior recognition task under thermal imaging video and the challenges and then describe the behavior recognition algorithm based on skeleton keypoints. In Section 3, the approach of BPF is introduced in detail. In Section 4, we build the dataset and analyze the effectiveness and robustness of our approach by simulating various behaviors of the elderly living alone indoors as an experimental application scenario. Finally, we conclude the paper and propose some possible future directions in Section 5.

2. Related Work

2.1. Thermal Imaging Behavior Recognition and Challenges

Behavior recognition is usually implemented based on wearable devices, environmental sensors, and vision. Wearable devices require the user to wear the device all the time,

requiring consideration of wearing comfort and device longevity. Deploying environmental sensors identifies behavior by detecting changes in physical signals, but this approach is heavily influenced by where the sensors are deployed. In contrast, it makes more sense to use visual information for surveillance. Still, visible light cameras inevitably involve an invasion of privacy in certain situations, which can easily make people resist. The thermal imaging camera only allows 8–14 μm infrared self-luminous irradiation to the camera sensor and thus imaging, not affected by external lighting conditions; therefore, it has better privacy protection. With the increasing awareness of privacy protection and the importance of 24-h surveillance, behavioral recognition methods for thermal imaging video [20–22] have been gradually proposed. All of the literature above migrated from visible video feature extraction methods to thermal imaging behavior recognition. The network recognition accuracy rises as the network structure deepens and becomes more complex.

The imaging of the thermal imaging camera is mainly dependent on the object's temperature but is also affected by the nature of the object, such as reflectivity. The expression of features, such as color, texture, and shadows in thermal imaging images, is not as good as visible images, resulting in increased difficulty in detecting and tracking human targets in thermal imaging images. In addition, the manufacturing difficulty of thermal imaging cameras and their relatively high price make the resolution of thermal imaging cameras at the civilian end generally low. Therefore, thermal imaging behavior recognition mainly has strong interference, few features, and low resolution.

2.2. Behavior Recognition Method Based on Skeletal Keypoints

The skeletal keypoint-based behavior recognition method utilizes pose estimation as a feature extractor, feeds the captured intrinsic dependencies of joints into a graph convolutional network promoted by a conventional CNN, and then classifies the action sequences of joint points by graph convolution. The classical pose estimation has two methods: bottom-up and top-down. The bottom-up approach is divided into two main processes: keypoints detection and keypoints clustering, as in OpenPose [23]. The top-down approach is divided into two main processes: target detection and single-person skeletal keypoints detection, as in AlphaPose [17]. The ST-GCN [14] can be considered as the pioneer of the behavior recognition method based on skeletal keypoints, and the subsequent literature [15,16] is based on this algorithm for improvement. Although such methods work extremely well in the visible light video, the reliance on the accuracy of the nodes obtained from pose estimation leads to the fact that they are still less than ideal for detecting small target humans in low-resolution images. In contrast, our BPF is explicitly proposed for small targets in low-resolution thermal imaging. By learning normal behavioral features, the model is more stable and easier to design.

3. Method

The pipeline of BPF is illustrated in Figure 2. The method is divided into three main modules: target detection, behavior parameter field establishment, and abnormal behavior recognition. The target detection module can accurately detect human targets using the improved cSE-YOLOv4 algorithm. The BPF module establishes a behavior parameter field for each human body to learn the normal behavior features of the human body. The abnormal behavior recognition module, using the feature matching method based on the intersection and ratio, analyzes the features to be measured of the human body and the behavioral features calculated in the BPF under the plurality of a certain position using cosine similarity and completes the detection of normal behavior as well as the two abnormal behaviors of fall and long-time immobility.

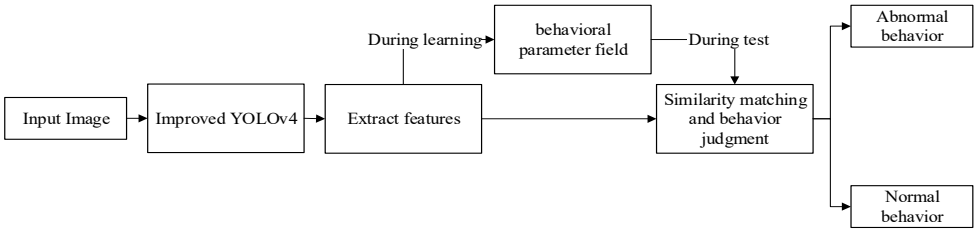


Figure 2. Overall structure block diagram.

3.1. Target Detection

To detect human targets quickly and effectively, we choose the YOLOv4 algorithm, a single-stage method with excellent performance, and improve it for small thermal imaging target scenarios.

3.1.1. YOLOv4 Introduction

The YOLOv4 is more complex in terms of network structure compared to the previous YOLOv3 [24], introducing the Mosaic data enhancement method and using the Mish activation function. The backbone network is CSPDarknet53, and the neck is a path aggregation network (PAN) with the addition of a spatial pyramid pooling (SPP) layer, which increases the perceptual field and makes it easier to separate the most significant contextual features, and the head still follows the YOLO-Head of YOLOv3, as shown in Figure 3.

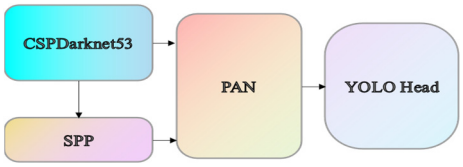


Figure 3. YOLOv4 network structure.

3.1.2. Improved YOLOv4 Algorithm

The scSE [19] module has shown excellent performance in semantic segmentation tasks. There are three variants of the scSE module, namely, the spatially compressed channel excitation module (sSE), the channel-compressed spatial excitation module (cSE), and the scSE module formed by combining the sSE and cSE modules.

1. sSE model;

As shown in Figure 4, the sSE model introduces the attention mechanism from spatial relations. The model activates the feature map of $H \times W \times C$ with $1 \times 1 \times 1$ convolution and the Sigmoid function ($\sigma(\cdot)$) to obtain the feature map of $1 \times H \times W$ and then multiplies it with the original feature map U space to obtain the new feature map $\hat{U}$.

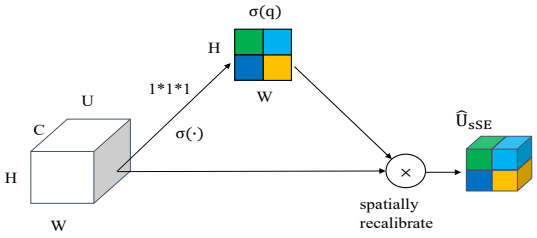


Figure 4. The structure of sSE.

2. cSE model;

As shown in Figure 5, the cSE model introduces an attention mechanism from the perspective of channel features to improve the network's ability to characterize important

channel features by compressing the global spatial features of the feature map. The process is to embed the input feature map $U = [u_1, u_2, \ldots, u_c]$ (where channel $u_k \in R^{H \times W}$), H and W are the height and width of the feature map, respectively) into the vector z (where $z \in R^{1 \times 1 \times C}$), C is the number of channels) through the global pooling layer with the kth channel out value as

$$z_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} u_k(i, j), \tag{1}$$

where i and j denote each position ($i \in H$, $j \in W$) on the kth channel feature map in the input feature map, respectively.

After two more fully connected layers and ReLU activation function ($\delta(\cdot)$) and Sigmoid function ($\sigma(\cdot)$), the importance degree $\sigma(\hat{z})$ of the channel is obtained as

$$\hat{z} = W_1(\delta(W_2 z)), \tag{2}$$

where W_1 and W_2 are the weights of the two fully connected layers.

3. scSE model;

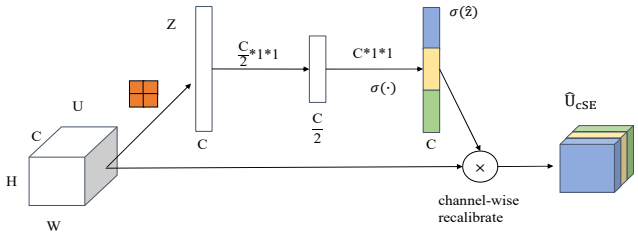
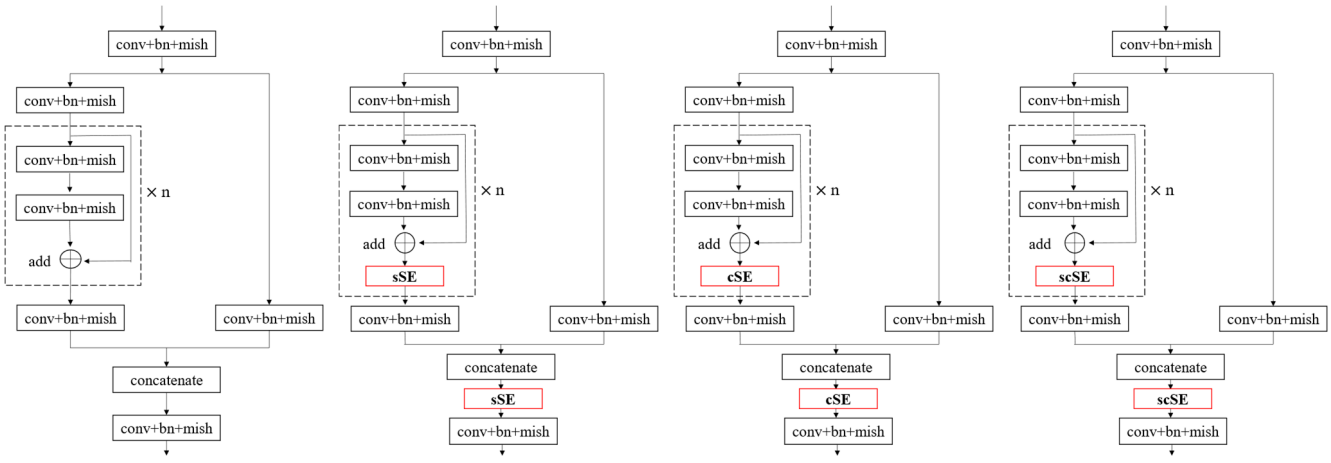


Figure 5. The structure of cSE.

The scSE module embedding in small target detection tasks helps to enhance the feature representation, which is also confirmed in the literature [25]. However, there is no set of theoretical statements on whether it is effective in small target detection tasks for thermal imaging spectroscopy.

In this paper, the scSE module is embedded into the backbone area of YOLOv4. Figure 6 illustrates the original ResBlock body as well as the three cases of embedded scSE modules.



(**a**) YOLOv4 (**b**) sSE-YOLOv4 (**c**) cSE-YOLOv4 (**d**) scSE-YOLOv4

Figure 6. (**a**) represents the structure of the ResBlock body in the original YOLOv4 backbone network; (**b**–**d**) represent the sSE module, cSE module, and scSE module embedded into the YOLOv4 backbone network, respectively.

3.2. Behavioral Parameter Field

Due to the limited pixel area occupied by human targets in low-resolution thermal imaging video, the pose estimation method extracts the human skeletal joints with false detection. Therefore, behavior recognition methods based on skeletal keypoints are not suitable for this task. From the perspective that the daily human behaviors in the indoor environment have regularity, this paper proposes a behavior recognition method called the behavior parameter field. The specific process of establishing the behavioral parameter field is shown in Figure 7.

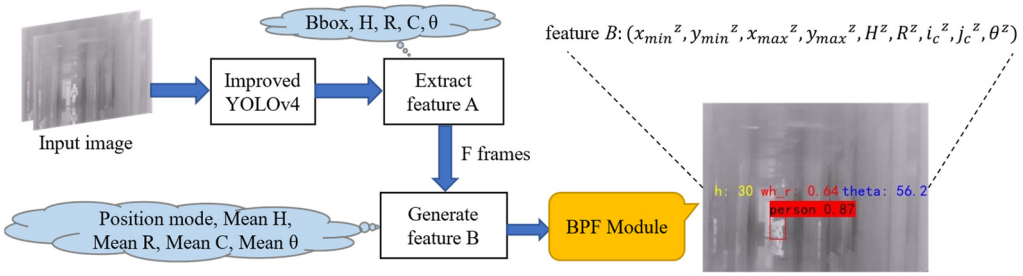


Figure 7. Flow chart for establishment behavior parameter field.

3.2.1. Feature Selection

1. Bounding box and aspect ratio;

With the YOLOv4 target detection model, the position information of the bounding box of the human target in the scene can be obtained: x_{min}, y_{min}, x_{max}, y_{max}. Further, the height and aspect ratio will change during human movement. Therefore, the width W, height H, and aspect ratio R of the target can be obtained by extracting the contour of the target as

$$R = \frac{W}{H}. \tag{3}$$

2. Center of mass;

The center of mass changes when people do different actions. In this paper, the center of mass is found for the ROI region of the human body found in the target detection stage using the method of calculating the center distance. For a two-dimensional discrete image $I(i,\ j)$, the $p+q$ order moments (m_{pq}) can be defined as

$$m_{pq} = \sum_{j=1}^{N} \sum_{i=1}^{N} i^p j^q I(i,j), \tag{4}$$

where p and q are non-negative integers. According to the above equation, the center of mass coordinates is (i_c, j_c) and $i_c = m_{10}/m_{00}$, $j_c = m_{01}/m_{00}$. It can be seen that the 0th and 1st order moments are the centers of mass.

3. Orientation angle;

According to the center moment obtained above, the orientation angle is calculated for the ROI area of the target detection, and the human body orientation angle can be found as

$$\theta = \frac{1}{2} arctan2(2m_{11}, m_{20} - m_{02}). \tag{5}$$

3.2.2. Behavioral Parameter Field Composition

The above features of the target in the thermal imaging video are collected over some time to initialize a history feature field for the target, which holds the normal behavior features of a human target for n consecutive frames. For the convenience of description, we denote the behavioral features of the target in the nth frame image as feature A, which is

$A = ({x_{min}}^n, {y_{min}}^n, {x_{max}}^n, {y_{max}}^n, H^n, R^n, {i_c}^n, {j_c}^n, \theta^n)$, where ${x_{min}}^n, {y_{min}}^n, {x_{max}}^n, {y_{max}}^n$ is the position region, H^n is the height, R^n is the aspect ratio, ${i_c}^n$ is the transverse coordinate of the center of mass, ${j_c}^n$ is the vertical coordinate of the center of mass, and θ^n is the direction angle.

Considering that the feature information of adjacent frames is relatively redundant, the first 4-dimensional features of F frames are accumulated to obtain the position plurality $({x_{min}}^z, {y_{min}}^z, {x_{max}}^z, {y_{max}}^z)$, and then to calculate the average of other dimensions. The final result of the calculation is called feature B, which represents the last feature of the human body at that position in the F frame, specifically $B = ({x_{min}}^z, {y_{min}}^z, {x_{max}}^z, {y_{max}}^z, H^z, R^z, {i_c}^z, {j_c}^z, \theta^z)$.

The calculation procedures for the mean value of height H^z, the mean value of aspect ratio R^z, the mean values of center of mass ${i_c}^z$ and ${j_c}^z$, and the mean value of orientation angle θ^z is as

$$\{H^z; R^z; {i_c}^z; {j_c}^z; \theta^z\} = \frac{1}{M}\sum_{n=1}^{M}\{H^n; R^n; {i_c}^n; {j_c}^n; \theta^n\}, M \le F, \tag{6}$$

where H^n, R^n, ${i_c}^n$, ${j_c}^n$, θ^n are the height, aspect ratio, center-of-mass horizontal coordinate, center-of-mass vertical coordinate, and orientation angle of the target in the position plurality in the nth image frame, respectively, and a total of M frames are involved in the averaging. F frames the plurality M of the target position in the image satisfies $M \le F$. The continuous F-frame feature A of the historical feature field is calculated according to the above method to obtain feature B. The region consisting of each mean feature of the human body at each position is called the behavioral parameter field.

3.2.3. Self-Update Learning Strategy

The features in the BPF model increase with time, and it is important to design an effective strategy to achieve feature updates. In this work, we explore several update strategies.

1. Direct update strategy;

The simplest update strategy is to use the newly written feature B to replace the existing features in the behavioral parameter field when the first 4 dimensions of the newly written feature B contain the position information that duplicates the existing feature positions in the behavioral parameter field. However, this operation directly causes the behavior parameter field to lose the ability to record historical features.

2. Voting update strategy;

Another natural update strategy is to establish a feature voting mechanism. Specifically, when the features containing the first 4 dimensions of location information are duplicated, they are recorded using a queue structure with storage space N. When $N = 3$ or $N = 5$, the 4-dimensional features with non-location information are voted one by one, and the result with the most votes in each dimension is selected as the final retention structure. However, this strategy results in a direct loss of the linkage of the dimensions in the behavioral parameter field.

3. Similarity update strategy;

We design a self-update strategy similar to the voting update strategy. Still, the queue with storage space N to record N features is used with repeated location information. When N features are satisfied, Euclidean distance similarity is calculated for the first $N-1$ features of 4 dimensions of non-location information and the first 4 dimensions of the Nth feature, and the feature with the highest similarity to the Nth feature among the first $N-1$ features is selected for replacement, and the replaced feature $B_{replace}$ is calculated as

$$B_{replace} = mindist_{ed} B_{bbox}^N,\ B_{bbox}, \tag{7}$$

where B_{bbox}^{N} denotes the feature at the Nth bounding box position and the first $N-1$ features located at the bounding box position of B_{bbox}^{N}. We will validate the proposed update strategy in the experimental section.

3.3. Abnormal Behavior Recognition

Abnormal behavior recognition is a binary classification problem. Our strategy is to develop classification rules based on the behavioral parameter field of different human body actions at each position and then determine whether there is abnormal behavior in the thermal imaging video. The specific flow chart is shown in Figure 8.

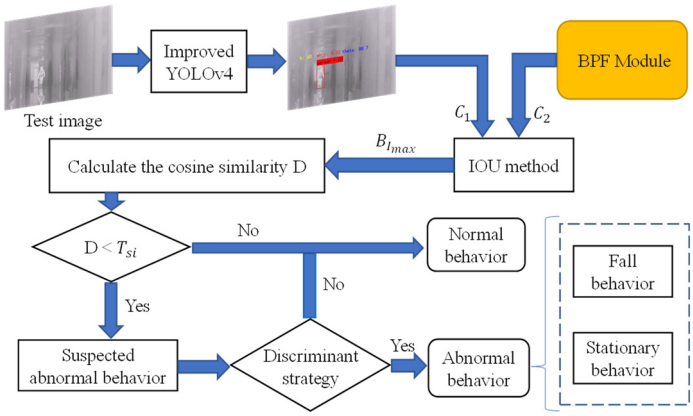


Figure 8. Abnormal behavior identification flowchart.

After cSE-YOLOv4 obtains the target bounding box of the current frame, this stage needs to match the target position of the current frame with the position in the BPF, and we propose the matching method based on intersection over union (IoU). The method is specified as follows: all the bounding boxes $({x_{min}}^z, {y_{min}}^z, {x_{max}}^z, {y_{max}}^z)$ in the behavior parameter field are taken as finite set C_1, and the bounding box obtained from the current frame is taken as finite set C_2, and the similarity between the two sets is calculated by the Jaccard coefficient. The one greater than the threshold $T_{iou2} \in [0,1]$ then gets the index I_{max} corresponding to the bounding box with the maximum similarity between the current frame and the behavioral parameter field in the selection list, so that the position matching between the current frame and the behavioral parameter field can be completed. The equation for the Jaccard coefficient as

$$J(c_1, c_2) = \frac{|c_1 \cap c_2|}{|c_1 \cup c_2|} \in [0,1]. \tag{8}$$

After location matching, define the current frame behavior feature as feature G, calculate the cosine similarity between feature G and feature B with index I_{max}, and get the cosine similarity D as

$$\mathrm{D} = \frac{\sum_{i=5}^{8} B_i G_i}{\sum_{i=5}^{8} B_i^2 \sum_{i=5}^{8} G_i^2}, \tag{9}$$

where B_i and G_i denote the components of the features B and G, respectively.

3.3.1. Judgment of Fall Events

The probability of a fall event is small and transient in nature. Therefore, when the above similarity match D is smaller than the similarity threshold T_{si}, the differentiation between the two behaviors for falling and lying on the ground is increased by introducing motion acceleration.

Taking the acceleration in the y-direction as an example, assume that the centers of mass of the target in the two consecutive images of the current frame are $({i_c}^n, {j_c}^n)$ and

$({i_c}^{n+1}, {j_c}^{n+1})$, respectively, and define the velocity of the moving target in the y-direction of the video in frame $n + 1$ as

$$v^{n+1} = \frac{{j_c}^{n+1} - {j_c}^{n}}{t}, \tag{10}$$

where t is the time between the two images, the acceleration of the elderly along the y-direction in the $n + 1$ image a^{n+1} is

$${a_y}^{n+1} = \frac{v^{n+1} - v^{n}}{t}. \tag{11}$$

Through the calculation of Equation (11), the acceleration of the moving object in the y-direction a_y can be found, and similarly, the acceleration in the x-direction a_x can be found.

For the fall event, if the above similarity matching D is less than the similarity threshold T_{si}, the acceleration in both directions is calculated separately for the moving target, and if a_x is greater than the judgment threshold T_{ax} or a_y is greater than the judgment threshold T_{ay}, then the fall event is considered to occur and Fall is set to true. The calculation formula is as

$$Fall = \begin{cases} true & if\left(a_x > T_{ax} \,||a_y > T_{ay}\right), \\ false & otherwise. \end{cases} \tag{12}$$

3.3.2. Judgment of Long-Time Immobility Events

For the prolonged immobility event, the prolonged immobility time counter C_{ina} is accumulated if the feature similarity of the two consecutive current images in frame F is greater than the threshold T_{ts} when the above similarity matching D is less than the similarity threshold T_{si}. Calculate the counter and the number of frames F its ratio; if it is greater than the judgment threshold $T_{ina} \in [0, 1]$, it is considered that there is a long inactivity time, and Ina is true to represent a long inactivity event. The specific discriminant criterion as Equation (13):

$$Ina = \begin{cases} true & if\left(\frac{C_{ina}}{F-1}\right) > T_{ina}, \\ false & otherwise. \end{cases} \tag{13}$$

4. Experiments

In the experimental phase, we take indoor elderly living alone as the experimental scenario and build the dataset by simulating various behaviors of indoor elderly living alone to analyze the effectiveness and robustness of the method in this paper.

4.1. Experimental Data

The only publicly available thermal imaging behavior dataset is the InfAR dataset from the literature [20]. Still, this dataset has a high image resolution and is not used for the small target problem. A thermal imaging dataset is constructed in this paper to study human abnormal behavior detection better. Due to the need to collect a large amount of data related to "falls" and the fact that older people often have mobility problems, we invited seven young volunteers to help with the filming. The dataset was filmed using a fixed thermal imaging camera with six categories and two different scenes. Each volunteer recorded four videos of each category in one scene, for a total of 336 videos. Each video is approximately 3 min long, with a resolution of 320 × 240 at 25 fps. The dataset contains normal behavior types (walk, sit down, bend over, and crouch down) and abnormal behavior types (stay still and fall), as shown in Figure 9.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 9. Self-collected dataset of thermal imaging abnormal behavior in indoor scenes. (**a**) walk (**b**) sit down (**c**) bend over (**d**) crouch down (**e**) fall down.

In the dataset partitioning section, we randomly selected 13 videos among the normal behaviors of each volunteer as the training set to establish the BPF model. The test set consists of two parts: first, five abnormal behaviors are randomly selected from the 16 abnormal behavior videos of the volunteer; second, three videos are selected from the normal behaviors of the volunteer that are not used as the training set, so the total test set is 56 videos.

4.2. Experimental Settings

Experimental Platform and Evaluation Metrics

The experiments were all conducted using the Pytorch 1.5.1 framework designed from Facebook to build the network, using an Intel Xeon(R)W-2133 processor at 3.60 GHz and an RTX2080 Ti GPU on a Windows 10 system. For the target detection part of the experiment, average precision (AP) and frames per second (FPS) are chosen as the algorithm measures. The behavioral detection section measures the results using accuracy (A_{cc}), false-negative rate (FNR), and false-positive rate (FPR). The Acc, FNR, and FPR are expressed as

$$\begin{cases} A_{cc} = \frac{T_P+T_N}{T_P+F_P+T_N+F_N}, \\ FNR = \frac{F_N}{T_N+F_N}, \\ FPR = \frac{F_P}{F_P+T_N}, \end{cases} \tag{14}$$

Assuming the existence of abnormal behavior of the elderly in a certain time period, if the algorithm successfully detects the abnormal behavior, it is called T_P, and if it fails to detect the abnormal behavior, it is called F_N. Assuming the absence of abnormal behavior in a certain time period, if the abnormal behavior is detected, it is called F_P, and if the abnormal behavior is not detected, it is called T_N.

4.3. Experimental Results and Analysis

4.3.1. Target Detection Network Training and Testing

1. Evaluation and analysis of the effectiveness of the improved YOLOv4 network model;

In this experiment, the added scSE module mentioned above is trained on the thermal imaging dataset established in this paper. The model performance is evaluated on the test set. The original YOLOv4 model is used as a baseline for performance comparison to study the performance changes brought by the three variants of SE embedded in different network positions. The experimental results are shown in Table 1.

Table 1. The performance evaluation of the three variants embedded in different positions of the network on the test set.

Model	AP@0.6/%	AP@0.7/%
YOLOv4(Baseline)	97.97	95.09
scSE-YOLOv4	95.67	92.63
sSE-YOLOv4	89.32	81.78
cSE-YOLOv4	98.31	98.31

The experimental parameters are set as follows: the input resolution of the network during training is 418 × 418, the network model is optimized by the Adam algorithm, the initial learning rate is set to 0.001, and a total of 100 epochs are trained. Table 1 shows the results of the model performance evaluation metrics obtained from the trained model tested on the test set, where the IOU threshold is set to 0.6 common to model evaluation, that is, AP@0.6. Similarly, AP@0.7/% represents the detection accuracy when the IOU threshold is 0.7.

It can be seen that embedding the cSE attention module after the "add" and "concat" feature fusion layer of the backbone can effectively improve the algorithm's performance in detecting small thermal imaging targets, with AP@0.6 and AP@0.7 values improving by 0.34% and 3.22%, respectively, compared to YOLOv4. However, embedding the scSE module and sSE module into the network's backbone results in different degrees of degradation. We attribute this to the fact that the cSE module enhances the learning of essential channel features in the thermal imaging images, but the addition of the sSE module compresses the channel features of the thermal imaging feature maps and disrupts the spatial representation of the features.

Figure 10 shows the comparison plots of the results of 4 consecutive frames obtained using YOLOv4 and the improved cSE-YOLOv4 algorithm in the detection of the self-collected thermal imaging dataset. It can be seen that the target frames predicted by the original network in frames 3 and 4 are not appropriate. In contrast, the improved network can give more accurate target frames, providing a more accurate human target frame for the next step of feature extraction of the behavioral parameter field.

Figure 10. The result of a comparison of four consecutive frames before and after improvement.

2. Performance comparison of different algorithms;

To verify the detection accuracy of different target detection algorithms for small targets under thermal imaging video, we conducted comparative experiments to test the performance of multiple methods on this thermal imaging dataset, and the results are shown in Table 2.

Table 2. Performance comparison of different algorithms.

Model	FPS	AP@0.6/%
SSD [26]	28	96.93
YOLOv3 [25]	23	97.28
YOLOv4 [18]	22	97.97
M2Det [27]	24	97.44
RFBNet [28]	27	93.98
cSE-YOLOv4	22	98.31

Combining Table 2 to compare the detection accuracy of different algorithms on IR datasets, the improved cSE-YOLOv4 model in this paper shows the highest detection performance and proves its effectiveness on small target detection.

4.3.2. Experimental Testing of Abnormal Behavior Detection

1. Comparison of experimental results of typical methods;

The behavior recognition method based on skeletal keypoints has excellent applications on visible data sets. However, compared to the rich and diverse datasets in the visible domain, there are no large datasets in the thermal imaging domain, so it is not reasonable to train the pose estimation algorithm specifically to obtain thermal imaging human keypoints. Additionally, the accuracy of skeletal keypoints is crucial for the final behavioral classification. In this paper, AlphaPose and OpenPose are selected as the algorithms for skeletal keypoints extraction to compare with the proposed BPF method, and thermal imaging images in two scenes are used as the input for the experiments. The experimental results are shown in Figure 11.

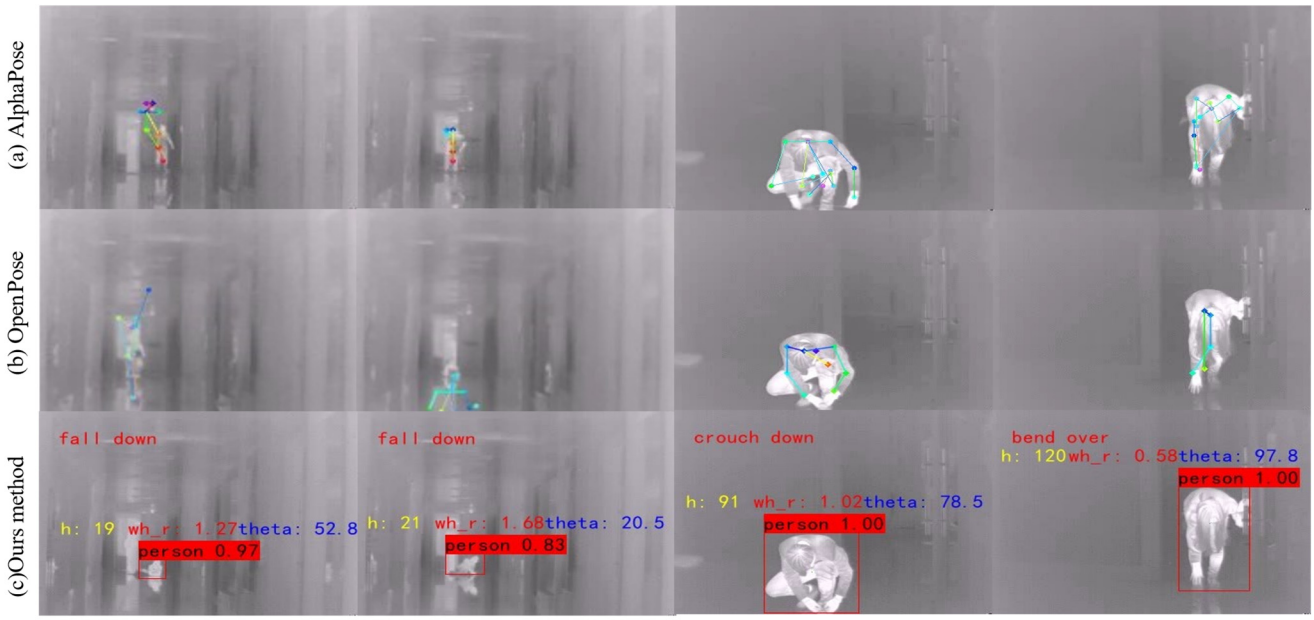


Figure 11. Comparison of results. (**a**,**b**) show that the extracted skeletal keypoints of AlphaPose and OpenPose in both scenes are not guaranteed to be accurate. Therefore, we can imagine the difficulty of the skeletal keypoint-based method for behavior recognition in this scene. Furthermore, (**c**) demonstrates that the method in this paper can still effectively extract the features required for the BPF (h is height, wh_r is the aspect ratio, and theta is orientation angle in (**c**)) and successfully detect the falling behavior.

2. Update strategy ablation experiment;

We test the impact of the behavioral parameter field established using different strategies on the performance of the two types of behavioral classification accuracy in the same scenario. The results are summarized in Table 3. We observe that similarity update strategies achieve better performance when the behavior parameter field is established for longer. This confirms the motivation of our design of this strategy which considers the relationship between the contact between different dimensions of feature B and the characteristics of the historical phase. Based on these observations, we use this similarity update strategy in the following experiments.

Table 3. Ablation experiments of three strategies.

Strategy	A_{cc} (Normal Behavior)/%	A_{cc} (Abnormal Behavior)/%
Direct update strategy	89.79	91.46
Voting update strategy	91.27	94.02
Similarity update strategy	91.42	94.28

3. BPF model performance test;

After testing, the detection results of our proposed method for identifying abnormal behavior based on the BPF model for thermal imaging video on the constructed test set are shown in Table 4. Experiments show that the method of human abnormal behavior recognition based on the BPF model can effectively detect the abnormal behavior in the video taken by a fixed thermal imaging camera, which has good practical value.

Table 4. Results of behavior detection.

Test Set	Number	A_{cc}/%	FNR/%	FPR/%
all	56	94.64	5.56	5
crouch down	5	80.00	0.00	20.00
sit down	7	85.71	0.00	14.29
bend over	3	100.00	0.00	0.00
walk	6	100.00	0.00	0.00
fall down	18	94.44	5.56	0.00
stay still	17	94.11	5.88	0.00

4. Sensitivity analysis of Hyperparameters;

In the process of establishing the BPF model, feature A needs to accumulate F frames to calculate feature B. By experimentally simulating the statistical results of elderly people's actions, the cumulative number of frames F is best when taken as 40. In updating the behavioral parameter field, N is best taken as 3 in the similarity update strategy. The parameters used in the experiments are: Jaccard coefficient threshold $T_{iou2} = 0.4$; normal behavior judgment threshold $T_{si} = 0.9$; horizontal acceleration judgment threshold $T_{ax} = 10$ and vertical acceleration judgment threshold $T_{ay} = 10$ in the fall event; and similarity threshold $T_{ts} = 0.95$ and prolonged immobility time judgment threshold $T_{ina} = 0.75$ in the prolonged immobility event. The values of the above parameters are obtained from extensive experimental verification and are used as the default parameters of the algorithm.

5. Conclusions

In this paper, a behavior recognition method is proposed based on BPF for the application scenario of a fixed thermal imaging camera. Taking the indoor scene of the elderly living alone as an example, the practicality and effectiveness of this method are verified. The method first detects human targets using the improved YOLOv4 target detection algorithm. Secondly, a BPF model is designed to characterize human abnormal behavior.

Then, the abnormal behavior recognition task is accomplished by a self-update learning strategy and a series of detection rules.

The work still has limitations, and future research can be conducted in the following aspects. (1) The current BPF method is mainly used for fixed thermal imaging camera scenes. The method can be extended to spherical thermal imaging cameras with rotatable viewpoints to establish a panoramic BPF model. (2) The features of the BPF model at this stage are extracted manually, and other effective features can be further explored in the future by considering using neural network learning to establish the BPF. (3) The idea of transfer learning can be applied to improve the skeleton extraction effect of the method based on skeletal keypoints on thermal imaging video, according to the unique imaging characteristics of thermal imaging. The skeleton and behavioral parameter field can be combined to obtain better recognition accuracy.

Author Contributions: Conceptualization, B.W., Z.D. and J.L.; Data curation, B.W., X.J. and J.L.; Formal analysis, B.W., X.J., Z.D. and J.L.; Funding acquisition, J.L.; Investigation, B.W. and J.L.; Methodology, B.W. and J.L.; Project administration, J.L.; Software, B.W.; Supervision, Z.D. and J.L.; Validation, B.W., X.J., Z.D. and J.L.; Visualization, B.W. and X.J.; Writing—original draft, B.W.; Writing—review & editing, B.W., X.J., Z.D. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Department of Science & Technology of Shandong Province (2017CXGC0810) and Shandong Natural Science Foundation (ZR2021QF043).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]
2. Bobick, A.; Davis, J. An appearance-based representation of action. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996; Volume 1, pp. 307–312.
3. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]
4. Shao, L.; Zhen, X.; Tao, D.; Li, X. Spatio-Temporal Laplacian Pyramid Coding for Action Recognition. *IEEE Trans. Cybern.* **2013**, *44*, 817–827. [CrossRef] [PubMed]
5. Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of local spatio-temporal features for action recognition. In Proceedings of the British Machine Vision Conference 2009, London, UK, 7–10 September 2009.
6. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.-L. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
7. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
8. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *arXiv* **2014**, arXiv:1406.2199.
9. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*; Springer: Cham, Switzerland, 2016; pp. 20–36.
10. Zhou, B.; Andonian, A.; Oliva, A.; Torralba, A. Temporal Relational Reasoning in Videos. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 831–846.
11. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
12. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
13. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
14. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
15. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7904–7913.
16. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 140–149.

17. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2353–2362.
18. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4 Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
19. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; pp. 421–429.
20. Gao, C.; Du, Y.; Liu, J.; Lv, J.; Yang, L.; Meng, D.; Hauptmann, A.G. InfAR dataset: Infrared action recognition at different times. *Neurocomputing* **2016**, *212*, 36–47. [CrossRef]
21. Wu, X.; Sun, S.; Li, J.; Li, D. Infrared behavior recognition based on spatio-temporal two-stream convolutional neural networks. *J. Appl. Opt.* **2018**, *39*, 743–750.
22. Akula, A.; Shah, A.K.; Ghosh, R. Deep learning approach for human action recognition in infrared images. *Cogn. Syst. Res.* **2018**, *50*, 146–154. [CrossRef]
23. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
24. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Jiang, R.; Peng, Y.; Xie, W.; Xie, G. Improved YOLOv4 Small Target Detection Algorithm with Embedded scSE Module. *J. Graph.* **2021**, *42*, 546–555.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
27. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9259–9266.
28. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 385–400.