IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2020, accepted January 28, 2020, date of publication February 3, 2020, date of current version February 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971283

Real-Time Detection and Motion Recognition of Human Moving Objects Based on Deep Learning and Multi-Scale Feature Fusion in Video

MEIMEI GONG AND YIMING SHU
School of Sports, Anhui Polytechnic University, Wuhu 241000, China
Corresponding author: Meimei Gong (gongmeimei@ahpu.edu.cn)

ABSTRACT At present, human body moving target detection and recognition algorithms based on deep learning have made breakthrough progress. However, in some applications with high real-time requirements, the existing deep learning real-time detection and recognition network is difficult to achieve high detection accuracy. Therefore, how to achieve accurate positioning and recognition of human moving targets while ensuring real-time detection is still an urgent problem in this field. Based on the single shot multi-box detector (SSD) real-time detection network, this paper proposes a real-time detection positioning and recognition network based on multi-scale feature fusion (IMFF-SSD), which improves the positioning accuracy and identification accuracy. First, this article analyzes the multi-scale features extracted from the SSD network. It combines the position-sensitive information provided by low-level detail features with the context information provided by high-level semantic features through feature fusion, which effectively improves positioning accuracy of the target prediction layer in the SSD network. Secondly, a feature embedded prediction structure is designed to strengthen the semantics of target features without changing the spatial resolution of the SSD prediction layer, and embed low-scale detailed features in high-semantic features for collaborative prediction of targets. This improves the accuracy of the SSD network's recognition of human moving targets at all scales. The experimental results show that by combining the above two improvements, the real-time monitoring and recognition network based on multi-scale feature fusion proposed in this paper has achieved a greater degree of improvement in positioning accuracy and motion recognition accuracy than the original SSD, which is better than some current the human body moving object detection and recognition algorithm has great advantages.

INDEX TERMS Deep learning, real-time, detection and motion recognition, multi-scale feature fusion.

I. INTRODUCTION

Today, surveillance cameras are used in all corners of our lives. The role of cameras is not just surveillance. It also helps us to obtain interesting areas and goals, and helps humans to better complete the expected work. In the field of machine vision, it has a very important role in target detection, recognition, positioning, tracking, and navigation [1], [2]. In the object detection of human motion, a large number of scholars have studied it, and many methods have been proposed to quickly and accurately find people in video images. However, due to the increased requirements and various needs, the goal of detecting people alone is not enough. In many scenarios, it is necessary to perform motion recognition on the detected people. Therefore, the real-time and accurate detection of the human body in the video image, and its positioning and motion analysis, have an important role in real life [3].

Video detection and recognition often encounter various problems under complex natural conditions [4]. Many human movement targets have their own gaps, and different categories have certain similarities. The video acquisition process is different due to the weather, environment, lighting, and shooting angle. There are also gaps in semantic understanding, as well as computational complexity and adaptability, which have brought great challenges to detection and recognition. At present, there are many models for monitoring and identification. The methods can be divided into: appearance-based models [5], motion-based models [6], space-time-based models [7], and deep learning-based methods [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Li.

Among them, deep learning is one of the focuses of current research. Convolutional neural network (CNN) is one of the deep learning models. It is a class of supervised machine learning algorithms that learn the hierarchical structure of features by constructing high-level features from low-level features. Cao *et al.* [9] proposed R-CNN (region-based convolutional neural networks). For the first time, deep learning was used for target detection. It is a target detection method combining convolutional neural networks and region prediction. The disadvantage of this method is that the size of the input image is consistent. To solve this problem, He *et al.* [10] proposed SPP-Net, which adds a spatial pyramid pooling layer to the convolution, which can transform features with different dimensions into features with the same dimensions and input fully connected layers. On the basis of R-CNN, Ren *et al.* [11] proposed faster R-CNN. Although the performance has been improved, they are all based on regional suggestion boxes, and the speed is still limited. In response to this problem, researchers have proposed a detection method without region suggestion boxes, originally YOLO proposed by Nguyen *et al.* [12]. This method has achieved the effect of real-time detection, but the detection effect is poor. Later, Chu *et al.* [13] proposed a single shot multi-box detector (SSD) model. This method improves the accuracy of the detection and also takes into account the speed of the detection. It is a relatively ideal algorithm. Both YOLO and SSD are end-to-end learning detection methods, which fundamentally solve the problem of detection speed [14].

At present, human body motion recognition mainly includes methods based on biomechanics [15], bioelectricity [16], and computer vision [17]. The methods based on biomechanics and bioelectricity mainly analyze motion by analyzing biomechanical information and bioelectrical information. Biomechanical information mainly includes joint angle, sole pressure, and so on [18]. These two methods require expensive information acquisition equipment and complicated deployment. The method based on computer vision is to use the camera to obtain the movement information of the human body, and then perform motion recognition and evaluation, such as Xie *et al.* [19] and Liu *et al.* [20] calculate the gait energy based on the contour information of the target in the video figure, and then realize motion recognition. The method based on computer vision is simple in equipment and convenient to deploy. It is the main method of motion recognition and evaluation at this stage. It is mainly divided into top-down and bottom-up detection methods [21]. The top-down detection method is to directly use the existing detector to estimate a single person's pose for each person in the image [22]. The detection time is directly proportional to the number of detections. As the target human body in the image increases, the detection time of each image also increases, and the bottom-up method [23], [24] can separate the target human body in the complex image. This method does not directly use the related information of other body parts and the global information of other people in the picture. However, the efficiency is not significantly improved, and the final local correlation requires a lot of time to calculate. Pishchulin *et al.* [25] proposed a bottom-up approach to associate some detection candidates with a single human body. However, the final detection time is relatively long. On the basis of Ref. [25], Insafutdinov et al. combined the detection method of image pairing scores with the ResNet network, which greatly improved the calculation efficiency [26]. However, it takes several minutes to detect each image, and real-time detection is still impossible. Ramanathan *et al.* [27] can obtain the motion energy map and the motion history map by superimposing the motion silhouette map of the human body. These two feature maps are matched with the template to achieve motion recognition. Hu *et al.* [28] used SIFT features to describe motion trajectories, and then used HMM models for motion recognition. These methods are relatively simple to calculate, but the recognition rate is relatively low. Shi and Luo [29] introduced the concept of ''entropy'' and proposed a human motion model based on motion energy, using dynamic time warping [30] algorithm to realize action recognition.

This article focuses on two basic problems in real-time human moving target detection networks, target positioning and recognition, and proposes a multi-scale feature fusion idea, which effectively improves the positioning accuracy and recognition accuracy rate of real-time human moving target detection networks (IMFF-SSD). First, this article analyzes the multi-scale features extracted from the SSD network. It combines the position-sensitive information provided by low-level detailed features with the context information provided by high-level semantic features through feature fusion, effectively improving the target prediction layer in the SSD network positioning accuracy. Secondly, a feature embedded prediction structure is designed to strengthen the semantics of target features without changing the spatial resolution of the SSD prediction layer, and embed low-scale detailed features in high-semantic features for collaborative prediction of targets. This improves the accuracy of the SSD network in identifying targets at various scales. The experimental results show that the network proposed in this paper has a greater degree of positioning accuracy and recognition accuracy than the original SSD.

Specifically, the technical contributions of our paper can be concluded as follows:

(1) This paper analyzes the multi-scale features extracted from SSD network, and combines the location sensitive information provided by low-level detailed features with the context information provided by high-level semantic features through feature fusion, effectively improving the positioning accuracy of the target prediction layer in SSD network.

(2) A feature embedded prediction structure is designed to enhance the semantics of target features without changing the spatial resolution of the SSD prediction layer, and embed low-scale detailed features in the high-semantic features for collaborative prediction of targets, thus improving the recognition accuracy of SSD network for targets of various scales.

The rest of our paper was organized as follows. Related work was introduced in Section II. Section III described the IMFF-SSD algorithm proposed in this paper. Experimental results and analysis were discussed in detail in Section IV. Finally, Section V concluded the whole paper.

II. RELATED WORKS

This section mainly introduces components commonly used to improve target detection and recognition performance, such as border regression, region generation, and non-maximum suppression, as well as several common evaluation indicators commonly used in target detection and recognition.

A. OBJECT DETECTION AND RECOGNITION COMPONENTS

The target detection and recognition algorithm based on convolutional neural network can automatically detect and recognize the target object in the image through self-learning. By adding some components to the detection algorithm, the detection performance of the entire network can be improved. For example, box regression, region generation, non-maximum suppression (NMS), and so on [31].

1) REGION GENERATION

Region generation refers to generating a series of rectangular borders on the input image, and each border contains potential objects that need to be identified, as shown in Figure 1. Two rectangular boxes surround the target. The red and yellow boxes in Figure 1 indicate the area where each target is located, and each frame contains potential objects that need to be identified.

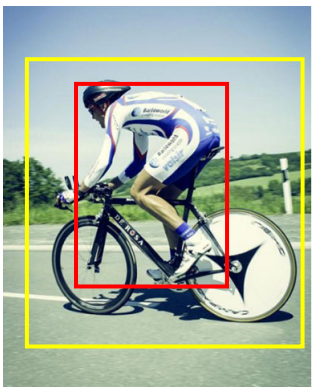


FIGURE 1. Region generation example.

In traditional object detection algorithms, the area generation algorithms are mainly divided into two categories [32]: one is a method based on super pixel merger; the other is a method based on edge information. Most of the methods based on super pixel merging combine some closely related super pixels to generate a general object area by a certain method. The area generation algorithm based on edge information makes full use of the intuitive characteristics of general objects to determine the possible areas of the object. However, these two types of methods are generally based on artificially designed features or classifiers. It is not easy to optimize through a large number of training samples, and it is also difficult to perform joint optimization with convolutional neural networks [33].

2) BORDER RETURN

For each area of interest in the image, not only can it be classified, but also the size of each area can be fine-tuned. As shown in Figure 2, when the range of the detection frame (solid line) is too large, the detection frame can be reduced to a suitable size (dashed line) according to the image content, and vice versa. This fine-tuning of the border size is called border regression. The red and yellow boxes in Figure 2 indicate the area where each target is located, and each frame contains potential objects that need to be identified.

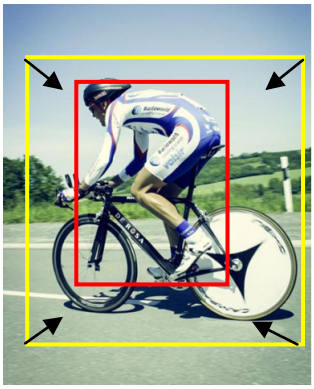


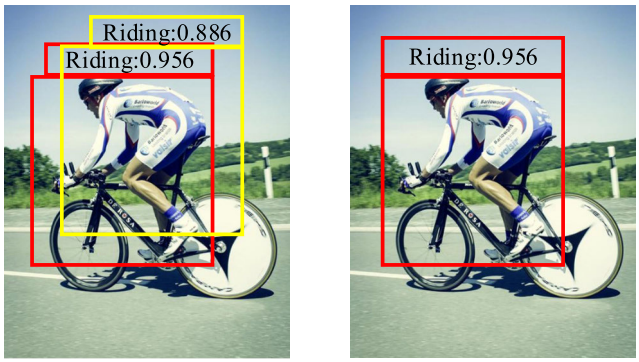
FIGURE 2. Border regression example.

Regression is often used in border regression to adjust the position of the detection frame. In order to make the regression target independent of the initial position of the detection frame, the offset of the regression frame is usually the offset from the solid line frame to the dotted frame in the Figure 2.

3) NON-MAXIMUM SUPPRESSION

When an object in an image is detected by an object detection algorithm, there may be multiple high-score detection frames around each target object. In order to remove redundant detection frames, each target object retains only one detection frame, and a non-maximum suppression algorithm needs to be used for subsequent processing. As shown in Figure 3, before the NMS, there were two detection boxes around the dog. One had a score of 0.956 in the decision box and the other was 0.886.After NMS, the low-score boxes with intersection over union (IoU) greater than the fixed threshold will be suppressed, leaving only the high-score boxes. Generally, the fixed threshold is set to 0.5.The calculation method of IoU is shown in equation (1).

$$IoU = area\,(b_1 \cap b_2)\,/area\,(b_1 \cup b_2) \tag{1}$$



(a) Before NMS (b) After NMS

FIGURE 3. Non-maximum fictional example.

Among them, $area\,(b_1 \cap b_2)$ represents the area of the intersection between the two detection frames b_1 and b_2; $area\,(b_1 \cup b_2)$ represents the area of the merged portion between the two detection frames b_1 and b_2.

YOLO, SSD, etc. used the idea of regression, and made great breakthroughs in detection speed. Although YOLO simplifies the entire object detection process and greatly improves the detection speed, reaching 45 frames per second, YOLO still has a lot of room for improvement. For example, no region generation is performed, and the regression can only be performed on 7×7 sub-regions to locate the target very accurately. SSD is based on a forward-propagating CNN network, which generates many fixed-size outer frames, and the possibility that each frame contains objects, and then performs a non-maximum suppression method to obtain the final predicted value [34]. After the VGG-16 network, the SSD adds convolutional feature maps with decreasing resolution layer by layer. These feature maps have different receptive fields, so the SSD can perform multi-scale object detection, that is, use high-resolution feature maps to detect images for small targets in the image, use low-resolution feature maps to detect large targets in the image. Therefore, based on the SSD real-time detection network, this paper proposes a real-time detection positioning and recognition network based on multi-scale feature fusion, which improves the positioning accuracy and recognition accuracy of targets at various scales.

B. PUBLIC EVALUATION INDEX

There are many evaluation indicators used to evaluate the performance of the target detection algorithm, such as precision, recall, average precision (AP), mean average precision (mAP), and detection speed.

1) ACCURACY AND RECALL

Suppose that in the target detection task, the test sample of i type is detected to obtain M_i detection frames, and $i \in \{1, 2, , C\}$. The symbol C is the total number of object categories to be detected. The M detection frames are sorted according to the confidence level from large to small, and the top N (top-N) detection frames are taken to calculate the correct rate and recall rate, and $N \in \{1, 2, , M\}$. When the IoU between the first N detection frames and the corresponding real frame is greater than a certain threshold, it is true positives, the number is recorded as tp_i^N, the rest are false positives, and the number is recorded as fp_i^N. The accuracy rate is actually the ratio of true positives in the top-N detection frames, as shown in equation (2).

$$P_i = tp_i^N/N = tp_i^N/(tp_i^N + fp_i^N) \tag{2}$$

Assume that there are K_i real frames in the type i test sample. The recall rate indicates that the ratio of true positives to the total number of real frames in the top-N detection frames is shown in equation (3).

$$R_i = tp_i/K_i \tag{3}$$

The accuracy rate indicator reflects the accuracy of the object detection algorithm to detect the object. The higher the accuracy rate, the lower the possibility of the detection algorithm's false detection. The recall rate indicator reflects the ability of the detection algorithm to detect objects. Generally, the higher the recall rate, the fewer objects the detection algorithm misses. Under a fixed IoU, the number of top-N detection frames is different, and the numbers of tp_i^N and fp_i^N are different, so P_i and R_i are different. Therefore, P_i and R_i can be regarded as a function of how many top-N detection frames are selected. Generally, the more the number of top-N detection frames is selected, the higher the recall rate will be, and the accuracy rate will generally decrease. Taking the recall rate as the abscissa and the correct rate as the ordinate, a precision-recall (P-R) curve can be obtained.

2) AVERAGE ACCURACY AND AVERAGE DETECTION ACCURACY

In addition to the performance of the detection algorithm represented by the P-R curve, the average correct rate AP is usually used to quantitatively measure the algorithm performance. In the target detection task, the AP evaluates the detection performance of a certain type of object by the detection algorithm, and the average detection accuracy mAP evaluates the detection performance of all types of object by the detection algorithm, that is, the average value of all types of APs.

3) DETECTION SPEED

The detection speed refers to the number of images that complete the target detection per unit time. The detection speed of the detection system is one of the important indicators to determine whether it meets the real-time application. Generally, the detection speed of the target is measured by the number of frames per second (FPS).

III. MULTI-SCALE FEATURE FUSION REAL-TIME MONITORING AND POSITIONING AND MOTION RECOGNITION NETWORK

A. SSD DETECTION NETWORK

The SSD network uses the VGG-16 network as its basic network. It changes its fully connected layers (fc6 and fc7) into

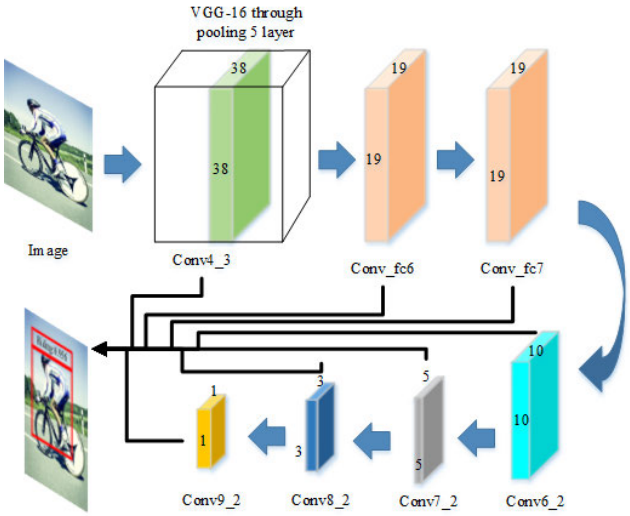


FIGURE 4. SSD network structure.

convolution layers (conv_fc6 and conv_fc7), and then adds convolution layers of different sizes to perform multi-scale targets prediction. Its network structure is shown in Figure 4. It can be seen from Figure 4 that several newly added convolution layers behind the VGG-16 network correspond to the characteristic response of the target at different scales. In addition to the conv4_3 feature layer of the VGG-16 network itself, conv4_3, conv_fc7, conv6_2 are used in total., conv7_2, conv8_2 and conv9_2 are six-layer feature layers to predict multi-scale targets. SSD networks are different from two-stage target detection networks, such as faster R-CNN, R-FCN, and so on. SSD network eliminates the generation of regional candidate frames, and uses full convolution to directly perform regression prediction on the position and category information of targets of different scales and shapes. At the same time, the predictions of targets at different scales are dispersed in different feature layers for implementation, thereby greatly improving the overall detection speed and accuracy of the network.

SSD target detection network combines the idea of sample space segmentation by anchor in faster R-CNN. Target prediction layers at different scales will generate default box shapes with different scales. These default candidate frames are centered on each pixel of the corresponding feature layer and are evenly distributed on the entire feature layer to predict targets of different scales and positions. The form of direct regression prediction of full convolution eliminates the process of generating regional candidate frames, which greatly improves the detection speed of the SSD network. And its structure is simple, the target prediction mechanism is single, and the detection accuracy is high. It can be used as the basic network of many networks. While maintaining its real-time detection performance, it can obtain higher detection accuracy. The next section will use the SSD network as the basic network. Through the multi-scale feature fusion structure and the Inception prediction structure, it can greatly improve the detection accuracy of the network while maintaining its high detection speed.

B. MULTI-SCALE FEATURE FUSION STRUCTURE

The SSD network implements hierarchical prediction of multi-scale human motion targets through feature layering. The lower-level feature layers are responsible for learning and predicting the features of smaller-scale targets, and the higher-level feature layers are responsible for learning and predicting large-scale target features. This article only uses each target prediction layer in the SSD network as the research object, and uses the small-scale detailed features learned by adjacent low-level feature layers as position-sensitive features, adjacent high-level semantic features as context information, and the corresponding target prediction layer features. Effectively combining them to effectively improve the accuracy of the SSD network's positioning of multi-scale targets. Considering the influence of the feature scale of each layer, only the three-level feature fusion is considered. Figure 5 shows the structure of an SSD network based on multi-level feature fusion.

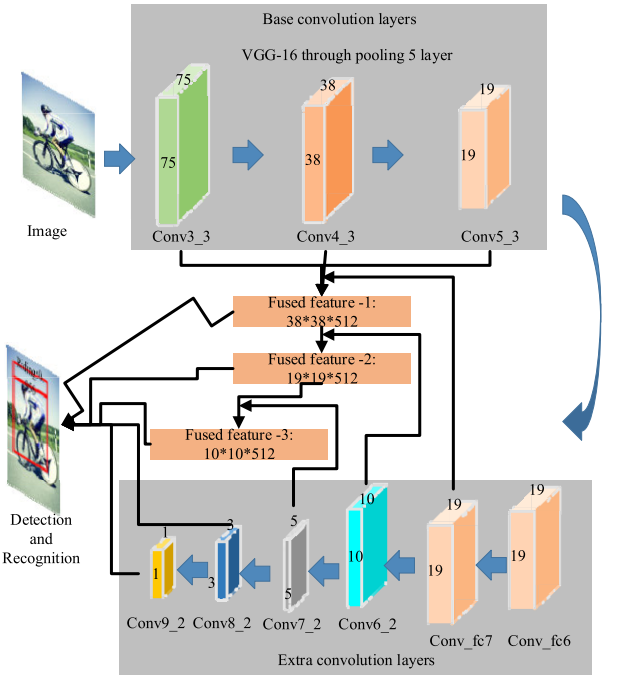


FIGURE 5. Structure of SSD network with multi-level feature fusion.

In order to effectively combine the extracted position-sensitive features and contextual feature information with target semantic features, this paper adopts a multi-scale human motion target feature-by-pixel corresponding addition method. In Figure 6, it is shown that two different features are fused in a pixel-by-pixel correspondence manner. To achieve the integration of two different features x and y in this way, it is necessary to ensure that the two feature dimensions and sizes remain the same before fusion. During the fusion process, the corresponding position elements of different

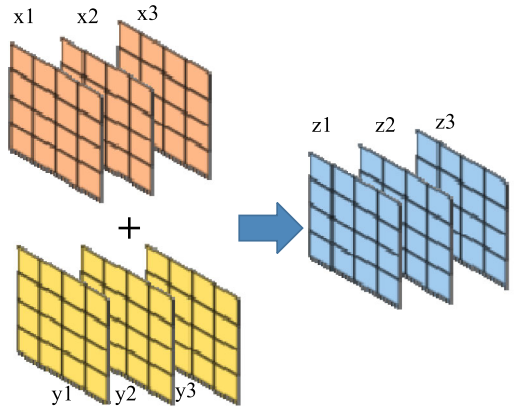


FIGURE 6. Feature fusion method based on pixel-by-pixel addition.

features are directly added, as shown in equation (4).

$$z_j^i = x_j^i + y_j^i \tag{4}$$

The feature dimension and size after fusion are consistent with the feature dimension and size before fusion.

In the SSD feature fusion network, the detailed features of the lower layer and the context information of the upper layer are added to the target prediction layer, and the semantic features of the target prediction layer still dominate the fusion features. Through the feature fusion method of corresponding element addition, the integrity of the fusion feature information is ensured, so that the target semantic feature can effectively select and combine the detailed feature and context information. The target feature itself provides a multi-scale feature fusion, this kind of constraint makes the fused features more effectively represent human moving targets.

Based on the above analysis, the fusion of multi-scale features by the pixel-by-pixel correspondence can effectively maintain the transfer of multi-scale information, and effectively combine the position-sensitive information provided by the detailed features with the context information provided by the high-level semantic features.

C. IMPROVED INCEPTIPN PREDICTION STRUCTURE

In traditional classification and recognition networks, such as VGG-16, ResNet-101, etc., the depth of the network is continuously increased to extract more abstract target features, thereby enhancing the recognition accuracy of the network. GoogleNet, on the other hand, has taken a different approach by using an Inception structure to increase the width of the network and extract target features through cascading. The Inception structure used in GoogleNet is shown in Figure 7.

Using the Inception structure to extract target feature information can greatly enrich the information density of each layer, thereby effectively making predictions on the target. In this paper, the concept of Inception structure combined with multi-scale target feature learning is used to solve the problem of insufficient semantics of each target prediction layer of the SSD network, thereby improving the accuracy

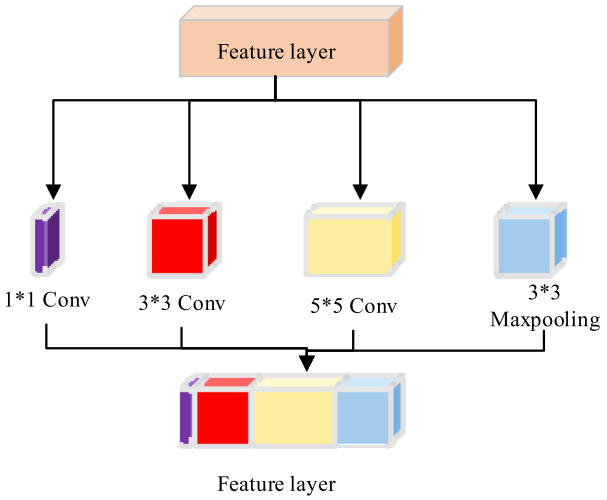


FIGURE 7. Inception structure diagram in GoogleNet network.

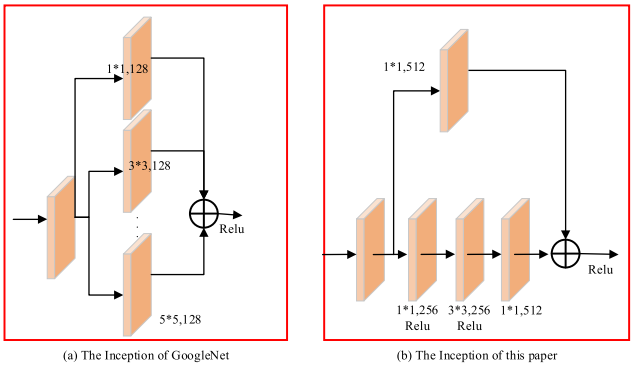


FIGURE 8. Improved inception prediction structure.

of target detection and positioning and motion recognition. However, such a high information density structure will greatly increase the computational cost. Therefore, this article makes certain improvements to the Inception structure in GoogleNet, making it more suitable for enhancing the feature semantics of each target prediction layer of an SSD network with a multi-scale feature fusion structure in order to achieve highly accurate recognition. The Inception prediction structure designed in this paper is shown in Figure 8. First, because the SSD network with a multi-scale feature fusion structure has realized hierarchical prediction of target objects of different scales, only the 1×1 and 3×3 two-way convolutions are retained in the Inception prediction structure for multi-scale feature extraction of targets. Secondly, in order to solve the computational cost problem of the Inception structure in the GoogleNet network due to the increase of the information density, this article uses the 1×1 convolution operation to control the feature dimensions in the designed Inception structure, thereby achieving more effective target features extract.

Therefore, the Inception prediction structure used in the SSD network for multi-scale feature fusion structure in this paper mainly includes two important branches:

the 3 × 3 convolution branch first uses the 1 × 1 convolution operation to reduce the dimension of the target feature. Then use 3 × 3 convolution to obtain the semantic enhancement of target features. Another 1 × 1 convolution branch realizes the extraction of lower-scale feature information of target features and combines it with features that are semantically enhanced through 3 × 3 convolution branches, so that the network can obtain higher semantic feature expressions at the same time, it can also obtain multi-scale feature expression of human moving targets.

D. MULTI-SCALE FEATURE FUSION REAL-TIME MONITORING AND POSITIONING AND MOTION RECOGNITION NETWORK

By adding the improved Inception structure to the target prediction layer of the SSD network with multi-scale feature fusion structure, it is used to improve the feature semantics of each target prediction layer of the network, thereby improving the accuracy of human body moving target detection and positioning.

Due to the multi-scale feature fusion structure of the SSD network, the conv4_3 feature convolution layer is mainly responsible for the learning of small-scale target features. Therefore, this article only adds this structure to the conv4_3 target prediction layer, and then extracts the target feature information. The specific structure of the SSD network with the Inception prediction structure is shown in Figure 9. After adding the Inception prediction module to the multi-scale feature fusion structure SSD network conv4_3 prediction layer, it is still connected to the 3 × 3 convolution layer for regression prediction of position and category information. The scale and the shape remains the same. This network, which is added to the Inception prediction structure after conv4_3 single layer for detection and positioning and motion recognition, is called IMFF-SSD.

The Inception structure combines the feature information of the human motion target extracted from these two channels, and uses this multi-scale information to predict the target collaboratively. The Inception structure improves the problem of insufficient semantics of target features in the target prediction layer. By selectively combining two features and cooperating with each other, multi-scale feature information can be obtained to represent human moving targets, thereby further enhancing the network the ability to detect and locate human motion targets and recognize motion.

During the training process, this article still uses the multi-task loss function of the original SSD to train the IMFF-SSD model. Equation (5) is the loss function corresponding to the IMFF-SSD network:

$$L_{x,c,l,g} = \frac{1}{N_{cls}} L_{cls}(x, c) + \frac{\varepsilon}{N_{reg}} L_{loc}(x, l, g) \tag{5}$$

In the expression, x is a binary vector whose constituent elements are $x_{ij}^{p} = \{0, 1\}$, which is used to indicate whether the default candidate box labeled i matches the target real box labeled j during training. Then the value of x_{ij}^{p} is 1,

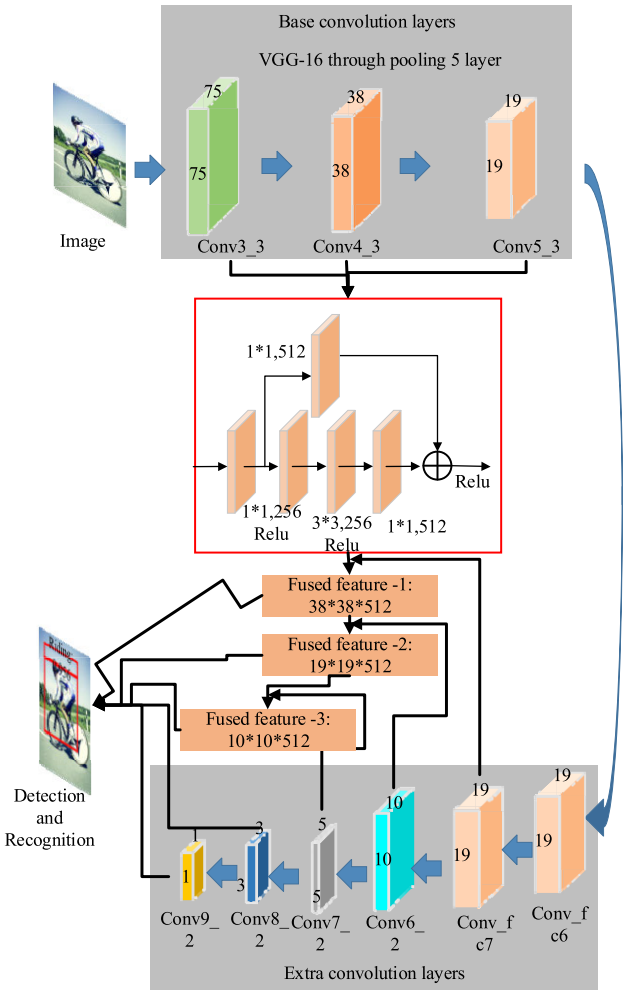


FIGURE 9. Network structure of IMFF-SSD.

indicating that the default candidate box is a positive sample, otherwise the value is 0, which is a negative sample. The superscript p indicates the category of the target that matches the default candidate box. Variable N_{cls} represents the total number of all positive and negative samples involved in the classification loss calculation, and N_{reg} represents the total number of positive samples. These two parameters are used to normalize the two items in the loss function. Variable α used as a balance factor to balance the contributions of two terms in the loss function. During training, the classification loss function $L_{cls}(x, c)$ uses the softmax loss function, and its specific expressions are shown in equation (6) and equation (7).

$$L_{cls}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log(\hat{c}_i^{p}) - \sum_{i \in Neg} \log(\hat{c}_i^{0}) \tag{6}$$

$$\hat{c}_i^{p} = \frac{e^{c_i^{p}}}{\sum_{p} e^{c_i^{p}}} \tag{7}$$

Among them, $\hat{c}_i^{p}$ indicates that the network scores the probability of the default candidate frame on the target category

Pos, *Pos* indicates the label set of all the default candidate frames that match the target real frame in a training batch, and *Neg* indicates that the label set of the default candidate frame that does not match the target true frame. The position regression loss function $L_{loc}(x, l, g)$ also uses the smooth L1 loss function proposed in fast R-CNN, and its specific expressions are shown in equation (8) to equation (12).

$$L_{loc}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} smooth_{L_1}(l_i^m - \hat{g}_j^m) \quad (8)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \quad (9)$$

$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \quad (10)$$

$$\hat{g}_j^w = \log(g_j^w/d_i^w) \quad (11)$$

$$\hat{g}_j^h = \log(g_j^h/d_i^h) \quad (12)$$

Among them, the variable $\hat{g}_j^m$ represents the relative position offset between the default candidate frame and the target true frame in (x, y, w, h) four coordinate dimensions. The variable l_i^m represents the predicted value of the four-dimensional coordinate information of the default candidate frame by the network. The variable (d_i^{cx}, d_i^{cy}) represents the coordinate information of the center position of the default candidate frame. The variables d_i^w and d_i^h correspond to the width and height of the i default box, respectively. The variable x_{ij}^p is only 1 when the default candidate box matches the target true box, indicating that the loss function is only involved in the calculation of the loss function when the default candidate box is a positive sample.

IV. EXPERIMENTS AND RESULTS

A. DATA SET AND IMAGE PREPROCESSING

This article uses the trained SSD model as a pre-trained model. All additional network layer parameters are initialized with ''xavier'', and the parameters are updated using the stochastic gradient descent algorithm. The initial learning rate is 10-3, and the weights are attenuated the term is 0.0005 and the momentum term is 0.9.

Camera was used for the experimental collection in this paper, and the image resolution was 1920 × 1080. The collected samples contained 10 people as the human motion targets to be detected. Finally, 952 valid samples were selected from all the samples taken, including samples from various angles. And by labeling all the samples manually, you need to label 10 targets in each sample, so there are a total of 9,520 target bounding boxes. The sample is divided into a training set and a test set, with a ratio of approximately 5: 1. And according to the data augmentation method of SSD network training data, the training data set is augmented with data. The data amplification methods mainly include the following categories:

Color space transformation: according to several standards for picture color change proposed in literature [35], the input training picture is transformed into color space.

Random cropping: use the original image as the training image to directly input the network for training; randomly sample a small block of the image so that it overlaps with the ground truth of the original target object is 0.1, 0.3, 0.5, 0.7, and 0.9. After cropping, the size of each cropped small image block is [0.1, 1] times the original image, and the aspect ratio is kept between 0.5 and 2. If the center of the target frame in the small image block falls within it, this paper keep the part of the small image block that overlaps with the original real target.

Random horizontal flip: the above-mentioned color space transformed and randomly cropped picture is flipped horizontally with a probability of 0.5, and the processed picture is scaled and deformed to a fixed size of 300 × 300.

B. MULTI-SCALE FEATURE FUSION STRUCTURE IMPROVES THE PERFORMANCE OF IMFF-SSD NETWORK

The analysis of the multi-scale feature extraction of the SSD network shows that the low-level layers of the network mainly extract the detailed features of human moving targets, the feature resolution is high, and there is a certain translational variability. The higher-level features have higher levels of abstraction, larger feature scales, and stronger semantics. They are suitable for robust feature extraction of human moving targets, and have certain translation invariance. In order to effectively improve the positioning accuracy of the network, IMFF-SSD combines low-level detail features with high-level semantic features and the target prediction layer, thereby improving the positioning accuracy of human motion targets. In order to further explore the sensitivity of the feature fusion network to the small offset of the target position, the detection target ''runner'' in the data set was shifted by 8 pixels relative position to compare the IMFF-SSD network's sensitivity to the target position offset sexual improvement.

As shown in Figure 10, the detection results of the SSD network and IMFF-SSD network before and after the target position is translated are shown.

In the two sets of comparison results of Figure 10(a) and Figure 10(b), the target in the lower image is shifted to the right by 8 pixels from the target in the upper image. The red box represents the target position detected by the corresponding network. The green box represents the position of the target true box. From the comparison of the two sets of detection results, it can be seen that the IMFF-SSD network effectively captures the small translation change of the target position, while the SSD network does not respond to the translation of the target position. Therefore, the IMFF-SSD network has a higher sensitivity to the position shift of the human moving target, which improves the positioning accuracy of the human moving target and can improve the performance of the IMFF-SSD network.

C. INCEPTION PREDICTION STRUCTURE IMPROVES THE PERFORMANCE OF IMFF-SSD NETWORK

In order to improve the feature semantics of each target prediction layer of the SSD network, so as to achieve more accurate recognition of human motion detection targets,

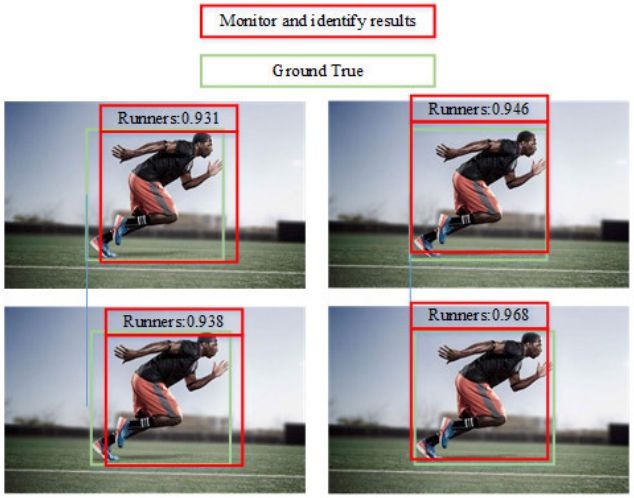


FIGURE 10. Comparison of the positioning accuracy of the SSD network and IMFF-SSD before and after the target is translated.

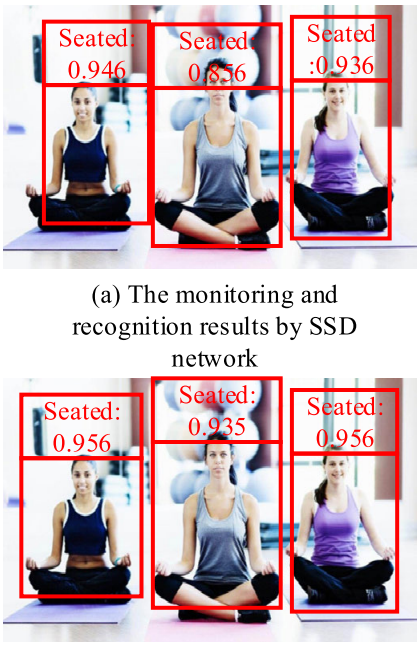


FIGURE 11. Comparison of monitoring results of human moving targets between SSD network and IMFF-SSD network.

the IMFF-SSD network uses the Inception prediction structure at each target prediction layer to predict multi-scale targets. In the case of changing the feature resolution level of the SSD network target prediction layer, the semantic improvement of multi-scale human motion target features is achieved, thereby effectively improving the overall recognition accuracy of the network. Figure 11 shows the comparison of the detection results of the same human moving target between the SSD network and the IMFF-SSD network. It can be seen from the comparison results that the detection accuracy of the IMFF-SSD network for the same detection of human moving targets is improved compared to the SSD network, and the detection accuracy of the IMFF-SSD network for human moving targets has reached 0.956. This shows that the IMFF-SSD network can extract higher semantic features of human moving targets. Therefore, the Inception prediction structure can improve the performance of IMFF-SSD networks.

D. IMFF-SDD NETWORK IMPROVES MOTION RECOGNITION PERFORMANCE

In order to realize the IMFF-SSD network to predict multi-scale human motion targets by using the Inception prediction structure at each target prediction layer, the semantics of multi-scale target features are realized without changing the feature resolution of the target prediction layer of the SSD network, thereby improving the performance of motion recognition. This article uses SSD network and IMFF-SSD network respectively, and evaluates the correctness rate, recall rate, average accuracy rate, and average detection accuracy for the data set. The experimental results are shown in Table 1.

From the comparison results, it can be seen that the four evaluation indexes of human motion recognition of the IMFF-SSD network are all improved compared to the SSD network. This shows that the IMFF-SSD network can extract higher semantic features of human moving targets, and realize highly accurate human motion recognition.

TABLE 1. Human motion recognition results.

Category		Seated	Standing	Pedestrian	Runner
SSD	Precision	0.971	0.951	0.861	0.767
	Recall	0.871	0.802	0.761	0.668
	AP	0.811	0.809	0.782	0.791
	mAP	0.798			
IMFF-SSD	Precision	0.978	0.963	0.889	0.839
	Recall	0.915	0.901	0.876	0.823
	AP	0.912	0.895	0.858	0.816
	mAP	0.870			

E. COMPARISON OF NETWORK DETECTION AND POSITIONING AND MOTION RECOGNITION PERFORMANCE

This paper conducts model training and testing of the IMFF-SSD network model, and compares the detection and positioning of the model on the test data set with the results of motion recognition with the current excellent algorithms in the correct rate, recall rate, average correct rate, and average detection accuracy. It shows the advantages of the IM-SSD network designed in this paper in detecting positioning

TABLE 2. Performance comparison between IMFF-SSD and various excellent human moving target detection and recognition networks.

Category		Seated	Standing	Pedestrian	Runner
R-CNN	Precision	0.899	0.821	0.751	0.763
	Recall	0.855	0.782	0.701	0.742
	AP	0.832	0.758	0.699	0.731
	mAP	0.755			
SPP-Net	Precision	0.859	0.815	0.764	0.763
	Recall	0.822	0.751	0.691	0.733
	AP	0.819	0.753	0.675	0.706
	mAP	0.738			
Faster R-CNN	Precision	0.912	0.895	0.815	0.796
	Recall	0.867	0.819	0.798	0.741
	AP	0.856	0.811	0.785	0.736
	mAP	0.797			
YOLO	Precision	0.961	0.951	0.859	0.798
	Recall	0.875	0.811	0.769	0.671
	AP	0.791	0.781	0.758	0.655
	mAP	0.746			
SSD	Precision	0.971	0.951	0.861	0.767
	Recall	0.871	0.802	0.761	0.668
	AP	0.811	0.809	0.782	0.791
	mAP	0.798			
IMFF-SSD	Precision	0.978	0.963	0.889	0.839
	Recall	0.915	0.901	0.876	0.823
	AP	0.912	0.895	0.858	0.816
	mAP	0.870			

and motion recognition performance. The results are shown in Table 2.

By comparing the results with other human moving target detection algorithms, it can be seen that the highest detection accuracy of IMFF-SSD single class reached 0.978, which is significantly better than some other excellent human moving target detection and recognition networks. In terms of overall detection performance, the IMFF-SSD network proposed in this paper has achieved relatively high detection results at this stage. From the comparison results in Table 2, it can be seen that IMFF-SSD network has more obvious improvement in detection and recognition of sitting, standing, and walking targets than SSD network.

Figure 12 shows the comparison of the P-R curve of the IMFF-SSD network and the comparison network. The larger the area under the curve in the P-R curve, the better the detection performance of the network. From the comparison results between IMFF-SSD and the comparison network in Figure 12, it can be seen that the P-R curve of the IMFF-SSD network is basically shown above the comparison network. From this, it can be seen that through comprehensive analysis, the IMFF-SSD network has more advantages and better detection accuracy than the comparison network.

F. COMPARISON OF NETWORK DETECTION SPEED

While ensuring the accuracy of the detection, the speed of the network and the real-time detection are also

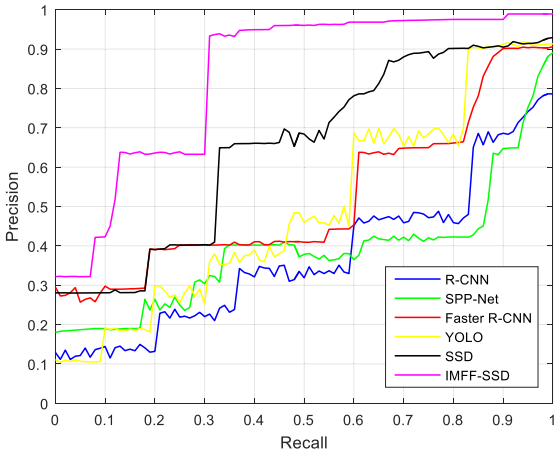


FIGURE 12. Comparison of P-R curve of IMFF-SSD and comparison network.

important factors. The IMFF-SSD model has reached a relatively high detection accuracy rate through the detection of the test data set. In order to explore its detection speed, this paper compares the test speed of IMFF-SSD with other end-to-end human moving target detection methods. The comparison results are shown in Table 3.

TABLE 3. Comparison of IMFF-SSD and other human moving target detection and recognition network speed results.

Method	mAP	FPS
R-CNN	0.755	85
SPP-Net	0.738	47
Faster R-CNN	0.797	79
YOLO	0.746	45
SSD	0.798	62
IMFF-SSD	0.870	43

During the test, this article sets the target confidence threshold to 0.05 to control the number of sample frames before performing non-maximum suppression operations, and then sets the overlap threshold to 0.3 to perform non-maximum suppression operations. From the comparison of the test speeds in Table 3, it can be seen that the detection speed of IMFF-SSD reaches 43 frames / second, and the optimal effect is achieved on the balance between the network detection accuracy rate and the detection speed. In addition, it can also maintain the original real-time detection performance of the SSD network, which opens up a large space for the promotion and application of the network.

V. CONCLUSION

With the development of science and technology, human body moving object detection methods based on deep learning have made great progress. The deep convolutional neural network trained based on massive data can extract more effective and robust feature information of the target, which greatly improves the accuracy of human motion detection

and positioning and motion recognition. Therefore, the performance of deep learning-based detection and recognition networks depends to a large extent on the network's ability to learn and predict target features. Based on the SSD network, this paper proposes a real-time detection location and recognition network based on multi-scale feature fusion. First, this article analyzes the multi-scale features extracted from the SSD network. It combines the position-sensitive information provided by low-level detail features with the context information provided by high-level semantic features through feature fusion, which effectively improves positioning accuracy of the target prediction layer in the SSD network. Secondly, a feature embedded prediction structure is designed to strengthen the semantics of human moving target features without changing the spatial resolution of the SSD prediction layer, and embed low-scale detailed features into human moving targets in high semantic features. Collaborative prediction is performed, thereby improving the accuracy of the SSD network in identifying human moving targets at various scales. The experimental results show that the network proposed in this paper has a greater degree of positioning accuracy and recognition accuracy than the original SSD.

REFERENCES

[1] V. Kaltsa, K. Avgerinakis, A. Briassouli, I. Kompatsiaris, and M. G. Strintzis, "Dynamic texture recognition and localization in machine vision for outdoor environments," *Comput. Ind.*, vol. 98, pp. 1–13, Jun. 2018.

[2] P. H. Song, J. H. Min, Y. S. Kim, S. Y. Jo, E. J. Kim, K. J. Lee, J. Lee, H. Sung, J. S. Moon, and D. H. Whang, "Rapid and accurate diagnosis of Clostridium difficile infection by real-time polymerase chain reaction," *Intestinal Res.*, vol. 16, no. 1, p. 109, 2018.

[3] Y. Wang and Y. F. Xu "Unsupervised learning of accurate camera pose and depth from video sequences with Kalman filter," *IEEE Access*, vol. 7, pp. 32796–32804, 2019.

[4] A. Aghraz, Q. Benameur, and T. Gervasi, "Antibacterial activity of *Cladanthus arabicus* and *Bubonium imbricatum* essential oils alone and in combination with conventional antibiotics against Enterobacteriaceae isolates," *Lett. Appl. Microbiol.*, vol. 67, no. 2, pp. 175–182, 2018.

[5] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, Jan. 2018.

[6] J. Ryding and I. A. Wernersson, "The role of reflection in family support social work and its possible promotion by a research-supported model," *J. Evidence-Based Social Work*, vol. 16, no. 3, pp. 322–345, 2019.

[7] A. M. Sánchez, P. Delgado, A. González-Rodríguez, C. González, A. F. G.-T. Rojas, and L. Lopez-Toledo, "Spatio-temporal approach for identification of critical conservation areas: A case study with two pine species from a threatened temperate forest in Mexico," *Biodiversity Conserv*, vol. 28, no. 7, pp. 1863–1883, Jun. 2019.

[8] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, and S.-F. Chang, "Multi-modal multi-scale deep learning for large-scale image annotation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1720–1731, Apr. 2019.

[9] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 836–846, Oct. 2017.

[10] X. He, C. Zhang, L. Zhang, and X. Li, "A—optimal projection for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1009–1015, May 2016.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[12] D. T. Nguyen, T. N. Nguyen, H. Kim, and H.-J. Lee, "A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 8, pp. 1861–1873, Aug. 2019.

[13] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion online hard example mining," *IEEE Access*, vol. 6, pp. 2169–3536, 2018.

[14] A. Moradi and S. M. Madani, "Technique for inrush current modelling of power transformers based on core saturation analysis," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 10, pp. 2317–2324, 2018.

[15] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.

[16] J. Huang, W. Huo, W. Xu, S. Mohammed, and Y. Amirat, "Control of upper-limb power-assist exoskeleton using a human-robot interface based on motion intention recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1257–1270, Oct. 2015.

[17] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, 2015.

[18] M. Gunaseelan, T. Kumaresan, R. Gandhinathan, and M. Ramu, "Biomechanical analysis of implantation of polyamide/hydroxyapatite shifted architecture porous scaffold in an injured femur bone," *Int. J. Biomed. Eng. Technol.*, vol. 30, no. 1, p. 16, 2019.

[19] S. Xie, X. Zhang, and J. Cai, "Video crowd detection and abnormal behavior model detection based on machine learning method," *Neural Comput. Appl.*, vol. 31, no. 1, pp. 175–184, 2019.

[20] G. Liu, S. Zhong, and T. Li, "Gait recognition method of temporal–spatial HOG features in critical separation of Fourier correction points," *Future Gener. Comput. Syst.*, vol. 94, pp. 11–15, May 2019.

[21] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Comput. Graph.*, vol. 85, pp. 15–22, Dec. 2019.

[22] E. Dogan, G. Eren, C. Wolf, E. Lombardi, and A. Baskurt, "Multi-view pose estimation with mixtures of parts and adaptive viewpoint selection," *IET Comput. Vis.*, vol. 12, no. 4, pp. 403–411, Jun. 2018.

[23] D. Prasanna and M. Prabhakar, "An effiecient human tracking system using Haar-like and hog feature extraction," *Cluster Comput.*, vol. 22, no. 2, pp. 2993–3000, 2019.

[24] M. Chen, S. Lu, Q. Liu, "Global regularity for a 2D model of electro-kinetic fluid in a bounded domain," *Acta Math. Appl. Sinica*, vol. 34, no. 2, pp. 398–403, 2018.

[25] L. Pishchulin, E. Insafutdinov, and S. Tang, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4929–4937.

[26] E. Insafutdinov, L. Pishchulin, and B. Andres, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 34–50.

[27] N. M. Thalmann, E. K. Teoh, M. Ramanathan, and W. Y. Yau, "Mutually reinforcing motion-pose framework for pose invariant action recognition," *Int. J. Biol. Macromolecules*, vol. 11, no. 2, p. 113, 2019.

[28] W. Hu, G. Tian, Y. Kang, C. Yuan, and S. Maybank, "Dual sticky hierarchical Dirichlet process hidden Markov model and its application to natural language description of motions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2355–2373, Oct. 2018.

[29] C. Shi and G. Luo, "A compact VLSI system for bio-inspired visual motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 4, pp. 1021–1036, Nov. 2018.

[30] A. Sharabiani, H. Darabi, S. Harford, E. Douzali, F. Karim, H. Johnson, and S. Chen, "Asymptotic dynamic time warping calculation with utilizing value repetition," *Knowl. Inf. Syst.*, vol. 57, no. 2, pp. 359–388, Nov. 2018.

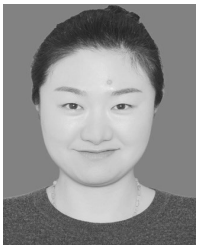
[31] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.

[32] R. Xin, Y. Yuan, and J. He, "High-efficient generation algorithm for large random active shield," *Sci. China Inf. Sci.*, vol. 62, no. 3, 2019, Art. no. 39108.

[33] W. Lin and Z. Hehe, "Application of Faster R-CNN model in vehicle detection," *J. Comput. Appl.*, vol. 38, no. 3, pp. 666–670, 2018.

[34] C. Li, D. Feng, Y. Hua, and F. Wang, "A high-performance and endurable SSD cache for parity-based RAID," *Frontiers Comput. Sci.*, vol. 13, no. 1, pp. 16–34, Feb. 2019.

[35] A. T. Deever and S. S. Hemami, "Lossless image compression with projection-based and adaptive reversible integer wavelet transforms," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 489–499, Jun. 2003.



MEIMEI GONG was born in Anhui, China, in 1980. She received the bachelor's degree from Xi'an Physical Education University, in 2002, and the master's degree from Nanjing Normal University, in 2009. She currently works with the School of Sports, Anhui Polytechnic University. She has published 11 articles. Her research interests include physical education and sports training.

YIMING SHU was born in Anhui, China, in 1986. She received the bachelor's degree from Wuhan Sports University, in 2008, and the master's degree from Anhui Polytechnic University, in 2011. She currently works with the School of Sports, Anhui Polytechnic University. She has published ten articles. Her research interests include sports news and physical education.