IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 14 January 2024, accepted 22 April 2024, date of publication 29 April 2024, date of current version 14 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394705

RESEARCH ARTICLE

A Novel Gesture Recognition Model Under Sports Scenarios Based on Kalman Filtering and YOLOv5 Algorithm

TINGTING WU[1] **AND XINGFENG FAN**[2]
[1]College of Physical Education, Xihua University, Chengdu 610039, China
[2]China University of Political Science and Law, Beijing 100091, China

Corresponding author: Xingfeng Fan (15848078221@163.com)

This work was supported by the Collaborative Education Project Between Industry and Academy under Grant 220605618160948.

ABSTRACT With the development of computer vision, automatic gesture recognition towards sports scenarios has been more significant in recent years. However, due to the complexity of motion scenes and the diversity of target objects, existing gesture recognition methods still face certain challenges in terms of fine-grained feature perception. To deal with this issue, this paper proposes a novel gesture recognition model under sports scenarios using Kalman filtering theory and YOLOv5 Algorithm. Firstly, the Kalman filtering algorithm is used to preprocess attitude data of targets. In particular, it can optimize the attitude data in time series by combining sensor measurements and system models. Thus, motion state decoding is completed for timed updates of trajectories. Then, the object detection algorithm YOLOv5 is introduced to detect gestures of humans. In this part, the initial YOLOv5 algorithm is lightly improved by introducing lightweight backbone structure, in order to improve both detection efficiency and running efficiency. Finally, the Kalman filtering part is combined with YOLOv5 algorithm part to construct a comprehensive gesture recognition model under sports scenarios. After that, some real-world images of sports scenarios are utilized as the experimental scene to testify performance of the proposed method. The results show that it has advantage in accuracy, stability, and real-time performance by comparing with typical models.

INDEX TERMS Kalman filtering, object detection, gesture recognition, computer vision.

I. INTRODUCTION

In recent years, the rapid development of intelligent computer vision technology has attracted widespread attention and research [1], [2]. Especially in sports scenes, accurate recognition and tracking of human posture is of great significance for various applications, such as sports training [3], [4], motion capture, medical rehabilitation, etc. However, due to the complex background, lighting changes, occlusion, and other factors in motion scenes, traditional pose recognition algorithms face enormous challenges [5]. To address this issue, we propose a new pose recognition model for motion scenes based on Kalman filtering and YOLOv5 algorithm. Kalman filtering, as a commonly used state estimation algorithm, can effectively handle the dynamic model and related measurement errors of moving objects. YOLOv5 is a fast and accurate object detection algorithm that can quickly locate and recognize key points in the human body.

This work aims to improve the accuracy and stability of pose recognition in motion scenarios by combining Kalman filtering with YOLOv5, fully utilizing their respective advantages [6]. Specifically, we first use the YOLOv5 algorithm to detect human key points in motion scenes, and then use Kalman filtering to track and predict the detection results to improve the accuracy and robustness of pose recognition [7]. We conducted extensive experimental evaluations on publicly available datasets and compared them with other commonly used pose recognition algorithms. The experimental results show that our proposed motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm has made significant progress in terms of accuracy and stability. Compared with traditional algorithms,

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

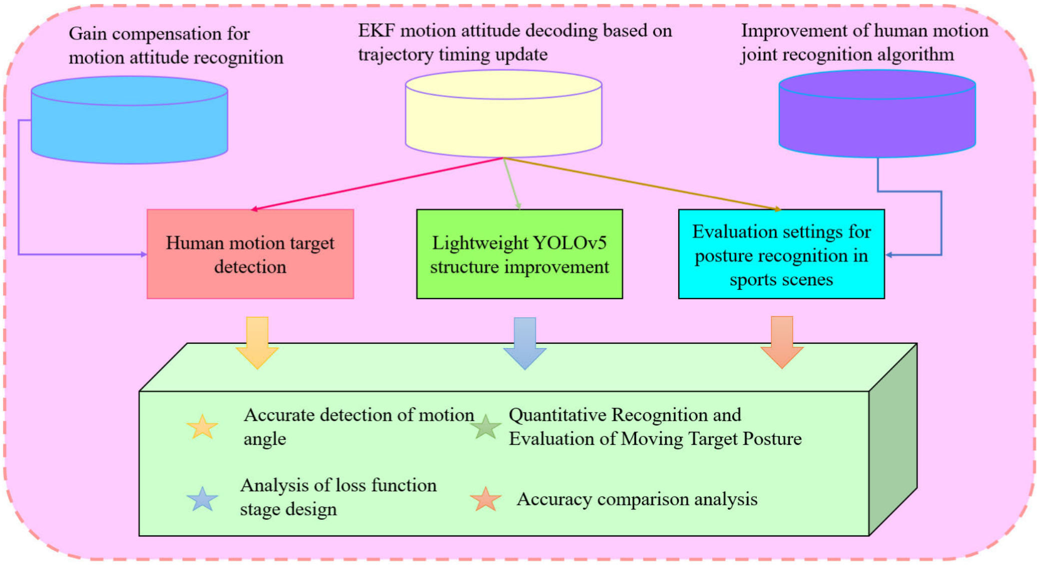
Gain compensation for motion attitude recognition
EKF motion attitude decoding based on trajectory timing update
Improvement of human motion joint recognition algorithm
Human motion target detection
Lightweight YOLOv5 structure improvement
Evaluation settings for posture recognition in sports scenes
Accurate detection of motion angle
Quantitative Recognition and Evaluation of Moving Target Posture
Analysis of loss function stage design
Accuracy comparison analysis

FIGURE 1. Main technical roadmap of this paper.

our model can better cope with issues such as changes in lighting, complex backgrounds, and occlusion, and has higher robustness and practicality.

Through the research in this article, we provide a new method and approach for solving pose recognition problems in motion scenes. We believe that this research achievement will have a positive impact on the development and application of intelligent computer vision technology, providing useful reference and reference for research and practical applications in related fields. The research process is shown in Figure 1. Main contributions of the research are summarized as follows:

1) By integrating the attitude tracking results of Kalman filtering with the attitude detection results of YOLOv5, a comprehensive model for motion scene attitude recognition was constructed, which can further improve the accuracy and robustness of attitude recognition

2) This study is based on the Kalman filtering theory and proposes an efficient attitude tracking method to address the potential blurring and noise issues that may occur during the motion process. By fusing the motion model of a moving object and sensor observation data, accurate tracking of the motion attitude is achieved, and real-time updates of attitude changes can be made.

3) This study uses the YOLOv5 algorithm as the main tool for attitude detection, which can detect the position and attitude information of moving objects in real-time under complex backgrounds.

Structure of the article is organized as follows. Section I discusses the current environment for motion scene recognition. Section II provides a more detailed explanation of previous research and the objectives of the article. Section III completed the pose calculation of Kalman filtering motion, performed gain compensation, and completed EKF decoding. In Section IV, the structural lightweight improvement of YOLOv5 for human motion object detection was completed, and a relevant evaluation system was established. Section V introduces the data sources used in this study and completes case validation. Section VI concludes this paper.

II. LITERATURE REVIEW

At present, there is relatively little research on motion scene pose recognition models based on Kalman filtering and YOLOv5 algorithm, but some scholars have already conducted relevant research in this area. Wang et al. [8] proposed a rope skipping posture recognition system based on Kalman filtering and YOLOv5 algorithm, which can perform real-time and accurate posture recognition of rope skipping movements. Daramouskas et al. [9] used a model based on Kalman filtering and YOLOv5 algorithm for skier pose recognition in motion scenes. They obtained pose data by placing sensors on the skier and combined it with image data to improve the accuracy and robustness of pose recognition.

DAHAL et al. [10] proposed a soccer player pose recognition method based on Kalman filtering and YOLOv5 algorithm. They used Kalman filtering to track key points of players and YOLOv5 algorithm to detect the position of players, achieving accurate recognition of soccer player poses. Golroudbari and Sabour [11] used a model based on Kalman filtering and YOLOv5 algorithm for basketball player pose estimation. They used YOLOv5 algorithm to detect basketball players and track key points of their bodies. At the same time, Kalman filtering was used to eliminate jitter and displacement errors caused during the motion process, ultimately achieving accurate pose estimation.

Chen and Hong [12] proposed a football player behavior analysis method based on Kalman filter and YOLOv5 algorithm. They used YOLOv5 algorithm to detect the position of football players, while using Kalman filter to track the movement trajectory of players, and combined action features for behavior analysis. They successfully achieved the recognition and analysis of football player behavior. The pedestrian pose recognition method proposed by Lin et al. [13] based on the Kalman filter and YOLOv5 algorithm is mainly divided into two steps. Firstly, the YOLOv5 algorithm is used to detect the pose, and then combined with the Kalman filter algorithm for pose smoothing. This method not only effectively recognizes the pedestrian pose, but also reduces pose errors and can handle noise and jitter in the video.

Li and Wu [14] used the YOLOv5 algorithm to detect targets in football scenes and smoothed the motion of the targets using the Kalman filtering algorithm, thereby achieving tracking and prediction of moving targets. This method not only accurately tracks moving targets in football scenes, but also responds to changes in scene and target motion speed. Yang et al. [15] used a YOLOv5 based pose recognition model and a Kalman filter based pose smoothing algorithm to recognize and estimate poses in motion scenes. This study adopted a new pose representation method to represent the direction and pose of objects in 3D space, thereby improving the accuracy and stability of pose estimation in motion scenes. At the same time, the research also combines the Kalman filtering algorithm with the state-of-the-art deep learning detector YOLOv5 to further improve the accuracy and stability of attitude estimation, achieving good performance.

Although these scholars' research work involves motion scene pose recognition or analysis based on Kalman filtering and YOLOv5 algorithm, each researcher will adopt different methods and model structures according to their specific field and purpose.

III. KALMAN FILTERING-BASED MOTION ATTITUDE DECODING

A. GAIN COMPENSATION-BASED GESTURE ATTITUDE PREPROCESSING

Gain compensation plays an important role in motion posture recognition, which refers to the recognition of human posture changes during motion through object detection and tracking from input images or videos. During this process, various factors such as noise, changes in lighting, and background complexity may cause errors and drift in attitude recognition. In an ideal sensor model, the three-axis cluster should share the same 3D orthogonal sensitivity axis that spans three-dimensional space, and the scaling factor converts the digital quantities measured by each sensor into actual physical quantities such as acceleration and gyroscopic rate [16]. However, practical applications are often affected by imprecise scaling, sensor axis misalignment, cross axis sensitivity, and non zero bias. Establish a three-axis accelerometer axis deviation framework [17].

$$s^B = Ts^S \tag{1}$$

$$\mathrm{T} = \begin{bmatrix} 1 & -\beta_{yz} & \beta_{zy} \\ \beta_{xz} & 1 & -\beta_{zx} \\ -\beta_{xy} & \beta_{yx} & 1 \end{bmatrix} \tag{2}$$

Among them, S^B and S^S represent specific accelerations, while Bij represents the rotation of the jth accelerometer around the j axis.

We have introduced the method of gain compensation. We compensate for the results of attitude recognition based on the weight adjustment strategy of Kalman filtering. When errors and drift occur in complex environments, we dynamically weighted average the detection results based on the weight adjustment strategy of Kalman filtering [18]. This can effectively suppress noise and interference, improve the stability and accuracy of attitude recognition [19]. The Kalman filter is used to estimate the state variable x ∈ Rn of sensors in a discrete time process, which is described by the following discrete stochastic difference equation:

$$x_k = F_k \cdot x_{k-1} + B_k \cdot u_{k-1} + w_{k-1} \tag{3}$$

We define the observation variable z ∈ R^m, where x_k represents the system state, F_k represents the state transition matrix, B_k represents the control input matrix, and u_{k-1} represents the control vector. The observation equation is obtained as follows:

$$z_k = H_k \cdot x_k + v_k \tag{4}$$

The random signals w_k and v_k represent the process noise and measurement noise of the sensor data, respectively. The sensor reading matrix is taken as H_k and, assuming they are independent of each other and conform to a normal distribution [20]:

$$\begin{cases} p(w) \sim N(0, Q) \\ p(v) \sim N(0, R) \end{cases} \tag{5}$$

where Q and R may vary with each iteration calculation. The Kalman filter estimates the process state by using feedback control, which estimates the state at a certain moment in the process and obtains feedback in the form of noisy measurement variables [21].

Gain compensation plays a crucial role in motion pose recognition. By introducing Kalman filtering combined with dynamic gain compensation strategies, we have achieved an efficient and accurate motion pose recognition model, which has important practical significance for many application fields such as motion analysis, sports training, and health monitoring.

B. KALMAN FILTERING-BASED GESTURE ATTITUDE DECODING FOR TIMED UPDATES OF TRAJECTORIES

Motion scene pose recognition refers to accurately estimating and tracking the pose or pose of an object from motion videos. In this study, I proposed a method based on Kalman Filter (EKF) to decode the pose of a moving scene. Kalman filter is a classic state estimation algorithm that estimates the motion state of an object through a statistical model and combines it with measured data for temporal updates [22]. EKF is a nonlinear filtering method that uses discrete difference equations to establish system state equations and measurement

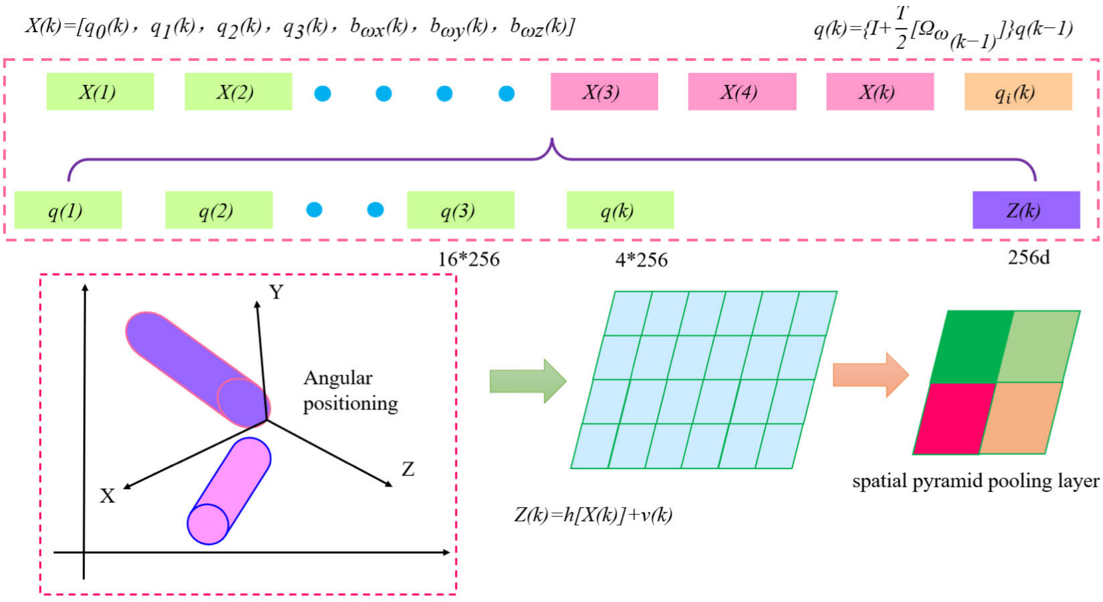
X(k)=[q0(k), q1(k), q2(k), q3(k), bωx(k), bωy(k), bωz(k)]
q(k)={I+T/2[Ωω(k−1)]}q(k−1)
X(1)
X(2)
X(3)
X(4)
X(k)
qi(k)
q(1)
q(2)
q(3)
q(k)
Z(k)
16*256
4*256
256d
Y
Angular positioning
Z
X
spatial pyramid pooling layer
Z(k)=h[X(k)]+v(k)

FIGURE 2. EKF-based motion attitude decoding.

equations, as follows:

$$\begin{cases} X(k) = f[X(k-1)] + G(k-1)w(k-1) \\ Z(k) = h[X(k)] + v(k) \end{cases} \tag{6}$$

Among them, $X(k)$ represents the state equation of the system, $z(k)$ represents the measurement equation of the system, *G(k-1)* is the driving matrix of the system noise, *w(k-l)* is the process noise of the system, and $v(k)$ is the measurement noise of the system.

The basic principle of a Kalman filter is to estimate the state of an object through a dynamic model, and update this estimation by measuring the model. In motion scene pose recognition, we use the position and motion state of the target object as the state vector, and the Kalman filter updates this state vector based on the observed position of the target object, as shown in Figure 2. We select the attitude quaternion of the streamer and the three-axis random drift error of the gyroscope as the state variables of the system, which are represented as:

$$X(k) = \left[q_0(k), q_1(k), q_2(k), q_3(k), b_{\omega x}(k), b_{\omega y}(k), b_{\omega z}(k)\right] \tag{7}$$

According to the Runge Kutta method, update the quaternion equation to:

$$q(k) = \{I + \frac{T}{2}[\Omega_{\omega_{(k-1)}}]\}q(k-1) \tag{8}$$

In order to decode the motion posture, we used trajectory timing update technology. Specifically, we decode the motion pose based on the output state vector of the Kalman filter and the pose information of the target object in the previous frame. By modeling the historical information of the target object's posture, we can more accurately predict and estimate the current posture of the target object. Kalman filtering is a recursive algorithm used to estimate the state of a system from a series of observations. In human motion joint recognition, I apply Kalman filtering to the process of predicting and updating joint positions. By combining predicted and observed values, Kalman filtering can correct joint position estimation and provide more accurate attitude information. In addition, I also introduced data association technology to solve occlusion and multi pose problems. Data association can help us match joint positions from different frames, thereby achieving tracking of joint positions in consecutive frames [23]. Through this method, even in complex motion scenes, human motion joints can be accurately identified.

When one limb is fixed and the other limb rotates arbitrarily around the joint center, the least squares method can also be used to estimate the vector from the joint center to the origin of the coordinate system on the rotating limb. Assuming that one limb is stationary and the other limb rotates arbitrarily around the joint center, the following equation holds:

$$\begin{aligned} &\left|\boldsymbol{a}_{2,t} - \boldsymbol{\Gamma}_{\boldsymbol{g}_2,\boldsymbol{o}_2}\right| - g = 0 \quad \forall t, \\ &\boldsymbol{\Gamma}_{\boldsymbol{g}_2,\boldsymbol{o}_2} = \boldsymbol{g}_{2,t} \times \left(\boldsymbol{g}_{2,t} \times \boldsymbol{o}_2\right) + \dot{\boldsymbol{g}}_{2,t} \times \boldsymbol{o}_2 \end{aligned} \tag{9}$$

where g is the magnitude of gravitational acceleration, which can be approximated as 9.8m/s^2, and $\boldsymbol{a}_{i,} - \boldsymbol{\Gamma}_{\boldsymbol{g}_i,\boldsymbol{o}_i}$ also represents the acceleration of gravity, so equation (9) holds. In the case where only one limb moves freely around the joint point, the Gauss Newton iterative algorithm can also be used to calculate the following optimization problem [24]:

$$\begin{aligned} &\min_{\boldsymbol{o}_2} \quad \sum\nolimits_{k=1}^{N} \boldsymbol{e}_k^2 \\ &s.t.\ \boldsymbol{e} \in \boldsymbol{R}^{N\times 1}, \boldsymbol{e}_k = \left|\boldsymbol{a}_{2,k} - \boldsymbol{\Gamma}_{g_2,a_2}\right| - g, k = 1, \cdots N \end{aligned} \tag{10}$$

By combining Kalman filtering and data association technology, we have improved the human motion joint

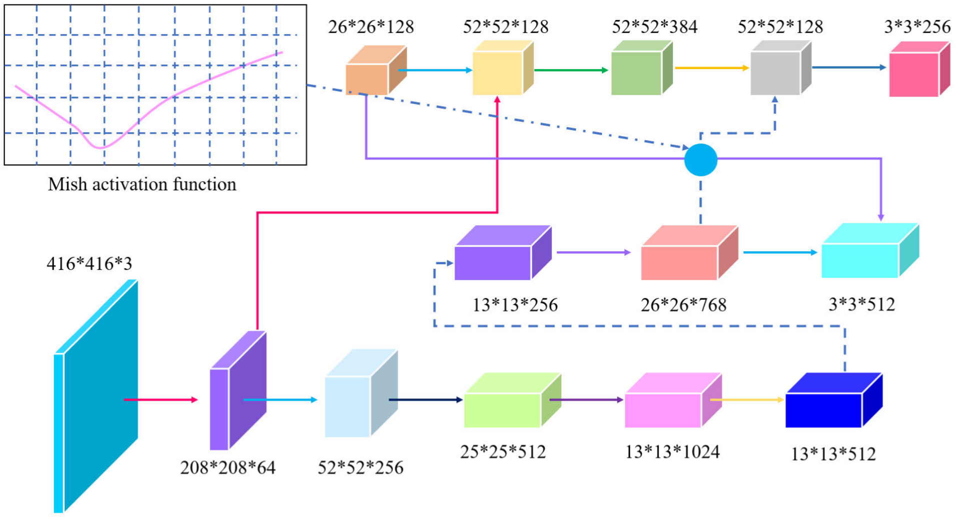

FIGURE 3. **Moving object recognition and detection.**

recognition algorithm, which may improve the accuracy and robustness of joint position. This will be verified in subsequent experiments. This study provides an effective solution for real-time motion tracking and posture evaluation.

IV. GESTURE RECOGNITION UNDER SPORTS SCENARIOS USING YOLOv5

A. BASIC YOLOv5 STRUCTURE FOR GESTURE RECOGNITION

When we perform human motion object detection, computer vision technology is usually used to identify human posture information in motion scenes. Regarding this issue, we will provide a detailed introduction to the application of YOLOv5 algorithm in human motion object detection tasks in the following text. This algorithm is currently one of the more advanced object detection algorithms based on deep learning [25]. The YOLOv5 algorithm can achieve real-time detection of various objects, and its anchor boxes technology can improve detection speed while ensuring detection accuracy. In the field of human motion object detection, we can achieve pose detection and tracking of the human body through a model based on YOLOv5, improving the accuracy and real-time performance of the model [26].

The image will generate redundant detection frames through the detector, and after passing through the single person human pose estimation algorithm, the redundant detection frames will also generate corresponding human pose representations, resulting in redundant pose estimation. At this point, it is necessary to construct Non Maximum Suppression (NMS) to eliminate redundancy. Set both Pi and P_j to represent human posture. c_i and c_j represent joint confidence levels for P_i and P_j postures, respectively. K_i and k_j represent the joint coordinates of the P_i and P_j postures, respectively. $B(kn\ i)$ is a bounding box centered on k_i, and its size is one tenth of the size of the bounding box B_i for the pose P_i. The joint point confidence elimination standard K_{sim} and joint point distance elimination standard H_{sim} are as follows:

$$K_{sim}(P_i, P_j|\sigma_1) = \begin{cases} \sum_n tanh\frac{c_i^n}{\sigma_1} \times tanh\frac{c_j^n}{\sigma_1}, & k_j^n \in B(k_i^n) \\ 0, & other \end{cases} \tag{11}$$

$$H_{sim}(P_i, P_j|\sigma_2) = \sum_n exp\left[-\frac{\left(k_i^n - k_j^n\right)^2}{\sigma_2}\right] \tag{12}$$

The confidence elimination criterion is to select the P_i with higher confidence as a reference between the attitude P_i and P_j to determine whether the attitude P_j has been eliminated. If the joint point k_j of attitude P_j is within the range of $B(k_i)$ and the confidence levels of the joint points k_i and k_j are similar, then the value of K_{sim} is 1, and in all other cases, the value of K_{sim} is 0.

The algorithms based on YOLOv5 have great application value in the field of human motion object detection. We can choose the appropriate algorithm based on actual needs and combine the two algorithms with graph convolution to improve the effectiveness and stability of the model. Moving object recognition and detection are shown in Figure 3.

In the propagation operation, each node in the graph will send its own feature information to its neighboring nodes. This propagation operation extracts and transforms the feature information of nodes to obtain useful information from their neighbors. Next, in the aggregation operation, each node in the graph will aggregate the feature information of its neighboring nodes. Through aggregation operations, nodes can fuse their local structural information with the feature information of neighboring nodes, thereby obtaining a more comprehensive feature representation. Finally, in nonlinear transformation processing, the graph convolutional algorithm uses nonlinear transformations to process the aggregated

TABLE 1. Setting of motion attitude angle recognition parameters.

Sports posture	Motion parameters	Angle range
Serve	knee joint	0-20
	ankle joint	50-90
	stretch	10-30
Backhand stroke	knee joint	10-30
	ankle joint	20-80
	stretch	10-30
Half squat	knee joint	0-90
	ankle joint	5-15
	stretch	20-40
Sprint	knee joint	0-60
	ankle joint	5-20
	stretch	10-30

feature information [27]. This nonlinear transformation processing can improve the representation ability of the model, thereby better capturing the complex relationships of nodes in the graph. It should be noted that in graph convolutional neural networks, local structural information is similar to the receptive domain in convolutional networks. They have the characteristic of sharing weights and are positively correlated with the number of layers in the network. The parameters used for extracting the motion angle are shown in Table 1.

B. IMPROVED YOLOv5 STRUCTURE BY INTRODUCING LIGHTWEIGHT BACKBONE

YOLOv5 is a popular real-time object detection algorithm that has significant advantages in accuracy and speed. In order to perform pose recognition in motion scenes, we need to make some improvements to YOLOv5. The main issue among them is the computational and parameter complexity of YOLOv5. To reduce computational load, we can achieve lightweight by reducing the number of layers in the model and the number of channels in the feature map. We can achieve this by removing certain convolutional layers and reducing the number of channels. However, it should be noted that while reducing model complexity, it is also important to ensure that the accuracy of the model is not significantly reduced. Therefore, careful trade-offs need to be made when lightweight YOLOv5 structure.

In the channel attention mechanism module, input a feature map F, assuming size C × H × W. Perform global maximum pooling and global average pooling on its size channels H and W, and then output two 1's × one × Process the feature map of C in a shared fully connected layer. In the shared fully connected layer, the first layer uses the ReLU activation function. The fully connected layer has C/r neurons, and the hyperparameter r represents the reduction rate. The number of neurons in the second layer is consistent with the channel C of the input feature map. Share and add the two values obtained from the fully connected layer, and then perform the Sigmoid activation function operation to ultimately generate the channel attention mechanism M_c module. Multiply the M_c module with the input feature map F element by element to generate the input features required for the spatial attention mechanism. The M_c module is shown in the formula:

$$M_c(F) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \tag{13}$$

Among, σ is the sigmoid activation function, $Fc\ max$ is the global maximum pooling based on channels H and W, $Fc\ avg$ is the global average pooling based on channels H and W, and W_0 and W_1 are weight parameters.

The introduction of attention mechanism stems from the fact that humans focus on global features with bright colors when processing image data, thereby neglecting local features. However, learning each local feature inevitably consumes time and effort. To avoid this problem, introducing an attention mechanism can strengthen the connection between local features and global features. Especially, the fusion of attention mechanism in convolutional neural networks can accelerate the calculation of data features, thereby simplifying the model [28].

$$M_s(F) = \sigma(f^{7\times7}_{Conv}([F^s_{avg}; F^s_{max}])) \tag{14}$$

where [;] represents the concatenation of feature maps, $f^{7\times7}_{Conv}$ indicates that the convolutional kernel is 7 × 7.

The CBMA attention mechanism focuses more on spatial features, while the SE attention mechanism only focuses on channel features. In the target detection task, combined with the relevant features of the dataset, the experimental results show that the fused CBMA attention mechanism has better performance than the SE attention mechanism. The CBMA attention mechanism can recognize more features and determine whether an object is valid by focusing on important information. By improving the lightweight YOLOv5 structure, combining Kalman filtering and prior knowledge, these improved methods can play an important role in building real-time attitude recognition systems and are widely applied in real scenarios. The improved network is shown in Figure 4.

C. SETTING FOR EXPERIMENTAL EVALUATION

In the field of motion scene pose recognition, the quality and scale of the dataset are crucial for the performance and robustness of the model. Therefore, it is necessary to evaluate indicators in this regard. For example, accuracy and recall can be considered to quantify the performance of the model. In addition, it is also possible to consider evaluating the generalization performance of the model, such as using methods such as cross validation and data augmentation to evaluate the gap between the model's performance and actual application scenarios [29].

Unlike in the field of object detection, where the intersection to union ratio (IOU) of detection boxes is used to measure the similarity between prediction boxes and real boxes, in human pose estimation, the indicator for quantifying the similarity between predicted and real values is OKS. The

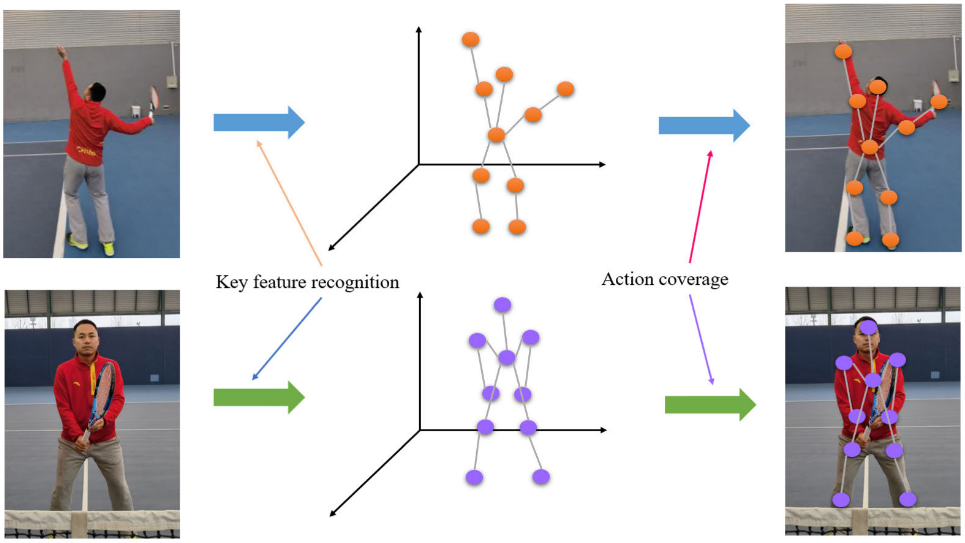

FIGURE 4. Lightweight YOLOv5 network structure.

specific calculation method is as follows:

$$OKS = \frac{\sum_i \left[exp\left(-\frac{d_i^2}{2s^2k_i^2} \right) \delta\left(v_i > 0\right) \right]}{\sum_i \left[\delta\left(v_i > 0\right)\right]} \tag{15}$$

where i is the index of the joint point. d_i is the Euclidean distance between the true and predicted coordinates of the i joint point. S is the size factor of the human body to which the joint belongs, and its value is the square root of the area of the human detection frame. K_i represents the normalization factor of the type to which the i joint belongs, representing the difficulty of annotation for that type of joint, and is a constant. V_i indicates whether the i joint point is visible. If it is not marked, then $v_i = 0$. If it is marked but occluded, then $v_i = 1$. If it is marked and visible in the image, then $v_i = 2$. If the conditions are met, then $\delta(\cdot) = 1$, otherwise $\delta(\cdot) = 0$, only calculate the joint points marked in the ground truth for cases where $v_i \geq 0$.

By setting the OKS threshold, it can be determined that the predicted value of key points greater than this threshold is a positive example. Taking a value every 0.05 from 0.5 to 0.95 as the OKS threshold, the accuracy AP of each threshold can be calculated using formula (16), and then taking the mean can obtain mAP [30].

$$AP = \frac{\sum_i \delta\left(OKS > s\right)}{N} \tag{16}$$

Among them, s represents the OKS threshold. I represents the index of all predicted joint points. N represents the total number of predicted joint points. When the OKS threshold is 0.5 and 0.75, the AP values are AP50 and AP75, respectively. The input image is at a medium scale (with an area range of 32 × 32 to 96 ×% The average accuracy between 6 pixels is APM, and the input image uses a large scale (area greater than 96) ×% The average accuracy of 6 pixels is APL.

In subsequent experimental verification, accuracy and recall are not the only evaluation indicators. For example, indicators such as real-time performance, scalability, and security also need to be considered. Therefore, when evaluating motion scene pose recognition models, in addition to considering algorithm performance indicators, it is also necessary to comprehensively evaluate the performance of the model in conjunction with practical application scenarios. The pseudo code of the proposed gesture recognition model under sports scenarios is shown in Algorithm 1.

V. EXPERIMENAL RESULTS AND ANALYSIS

We take the tennis sports as the object scenario, and collect some realistic images from our sports practice to specifically construct a dataset. We totally collect 480 images for dataset construction, and such dataset contains ten different gestures in tennis sports. In this section, the proposed recognition model and several typical comparison methods. The obtained verification results are obtained and analysed to demonstrate proper performance of the proposed method.

A. ACCURATE DETECTION OF GESTURE ANGLE

When it comes to motion scene pose recognition models, accurate detection of motion angles is an important research direction. Movement angle refers to the posture angle of an object or human body during movement, the joint angle of the human body, or the rotation angle of the object. In order to train and evaluate accurate motion angle detection models, an accurately annotated dataset is required. For human pose recognition, research is using human keypoint annotation to capture the position and angle of joints. For the detection of object rotation angle, it is necessary to accurately label the object's rotation angle. Ensuring the accuracy of dataset annotation is crucial. Use different methods to estimate the

Algorithm 1 Kalman Filtering and YOLOv5 Algorithm-Based Gesture Recognition Model Under Sports Scenarios

```
Input: Specific acceleration $s$ $S^B$ and $S^S$, the
state variable $x$ of the sensor in a discrete
time process, state Equation $X(k)$ of the
System, coefficient of variation $i$.
Calculate $S^B$ using formula 1 using $S^S$
Introduced the method of gain compensation
$i \in$ R
for all $i = 1$ to $R$ do
    Calculate $X_t$ using eq-3
    $x_k = F_k x_{k-1} + B_k u_{k-1} + w_{k-1}$
    Process noise and measurement noise in
    collecting sensor data
    for $x_k$ 1 : $i$
        Calculate $X(k)$ according to eq-7
        if the pose information of the target
        object in the previous frame is cluttered
            Modeling Historical Information of
            Target Object Posture
            Estimating the vector of the origin of
            the coordinate system on a rotating limb
        else
            Determine if the posture has been eliminated
    end for
end for
```

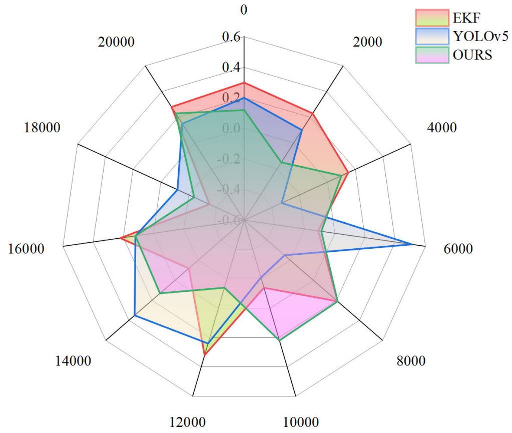

FIGURE 5. Motion angle detection results.

motion angle based on specific problems. For continuous angle values, use regression methods to predict angles. For discrete angle classification, classification methods can be used to predict the category of angles. Ensure that the model can accurately regress or classify motion angles. The results are shown in Figure 5.

The data in Figure 5 shows that at each time point, the operation time of different algorithms is given. Based on the data, it can be seen that the time points are given in increments from 0 seconds to 20000 seconds. According to the data, the EKF algorithm calculates motion recognition angles (in degrees) at different time points, ranging from 0.3 to -0.14. Among them, the positive and negative angles represent different directions of motion, for example, 0.3 represents a forward motion angle. The YOLOv5 algorithm calculates motion recognition angles (in degrees) at different time points, ranging from 0.2 to 0.51. Among them, the positive and negative angles represent different directions of motion, for example, 0.2 represents a forward motion angle. The algorithm obtained in this study calculates motion recognition angles (in degrees) at different time points, ranging from 0.12 to 0.23. Among them, the positive and negative angles represent different directions of motion, for example, 0.12 represents a positive angle of motion.

B. QUANTITATIVE RECOGNITION OF GESTURES IN MOVING TARGETS

When it comes to quantitative recognition and evaluation of the pose of moving targets, a motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm can provide an effective solution. The goal of posture quantification is to objectively evaluate posture by identifying and quantitatively describing the posture and actions of moving targets in the scene. In the quantitative recognition and evaluation of motion target posture, we collect image or video data from the motion scene and preprocess it, including image denoising, image enhancement, and other operations to improve the accuracy of subsequent processing. Using the YOLOv5 algorithm to detect targets in the preprocessed image, locate and recognize the position of moving targets. Based on the Kalman filtering algorithm, predict and estimate the position and velocity of the target, and extract attitude information. Based on the definition and requirements of the target posture, design appropriate evaluation indicators and algorithms to quantitatively evaluate the posture, such as angle, motion amplitude, etc. Based on the quantitative evaluation results, conduct data analysis and statistics, and provide corresponding attitude recognition evaluation reports, including accuracy, stability, and other related indicators. The results are shown in Figure 6 and Figure 7.

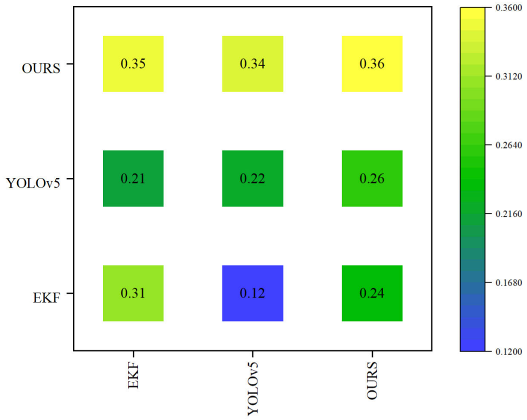

FIGURE 6. Comparison of quantitative evaluation combinations of different algorithms.

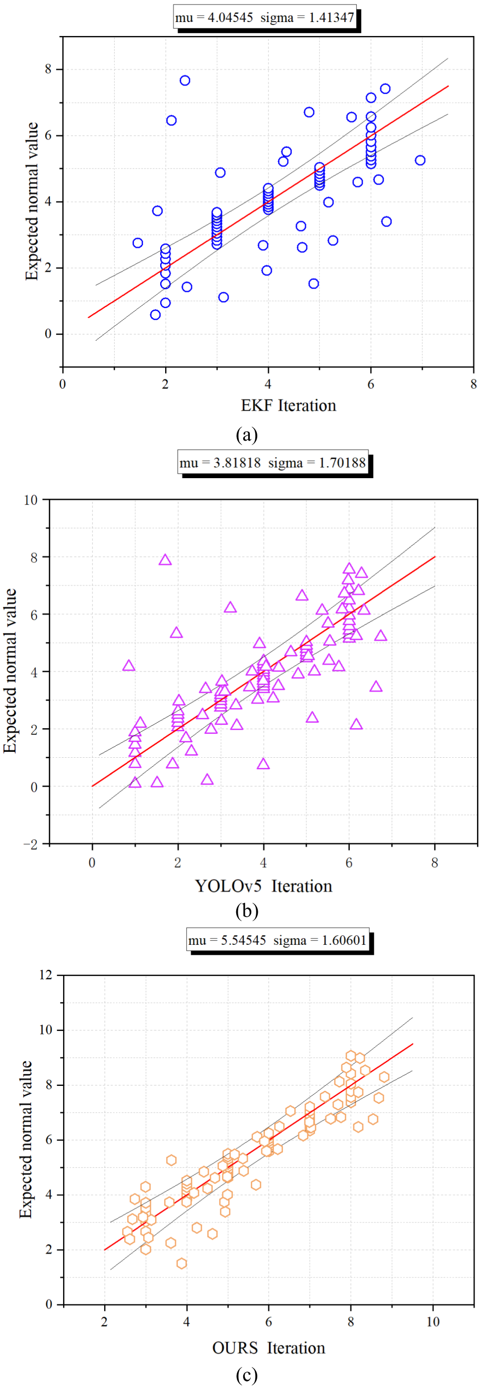

FIGURE 7. Distribution of evaluation results.

The results in Figures 6 and 7 show that the average evaluation score for EKF motion target pose quantification recognition is 3.75, with a maximum score of 6 and a minimum score of 2. The average score of YOLOv5 motion target pose quantification recognition evaluation is 3.62, with a maximum score of 6 and a minimum score of 1. The average score for quantitative recognition and evaluation of the posture of the OURS moving target is 5.15, with a maximum score of 8 and a minimum score of 3. The score fluctuation of the EKF model in quantitative recognition and evaluation of moving target posture is small, ranging from 2 to 6. This indicates that the performance of the model is relatively consistent and stable across different samples. The YOLOv5 model has the widest score range, ranging from 1 to 6. This indicates that the model exhibits significant differences in performance across different samples, which may be influenced by specific scenarios or data.

The score range of the model obtained in this study is also relatively wide, ranging from 3 to 8. However, the average score of the model is the highest, which may indicate that the model performs better on most samples. Meanwhile, this may also mean that the model performs exceptionally well on certain samples. The YOLOv5 model has certain fluctuations in scores and may require further tuning and improvement. The EKF model is relatively stable in scoring and performs average. The final selection should comprehensively consider the performance, stability, and applicability of the model. The computational model obtained in this study performs best in the quantitative recognition and evaluation of moving target posture, with the highest average score. Next is the EKF model, followed by the average score, and finally the YOLOv5 model, with the lowest average score.

C. ANALYSIS FOR STATUS DESIGN OF LOSS FUNCTION

The loss function plays a crucial role in attitude recognition algorithms, affecting model training and performance optimization. In the motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm, it is crucial to design a scientific and reasonable loss function. The following are some commonly used loss functions and their design analysis: This loss function is one of the most commonly used attitude estimation loss functions, which evaluates the accuracy of the model by calculating the sum of squared errors between the predicted value and the true value. In the design process of the loss function stage, the mathematical characteristics and application scenarios of the mean square error loss function can be combined to optimize and adjust the parameters in the model to achieve higher accuracy and accuracy. The results are shown in Figure 8.

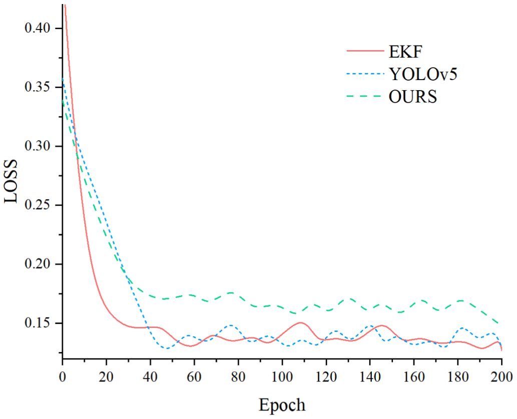

FIGURE 8. Different iterative calculations of loss functions.

Figure 8 shows that the loss function loss values of different algorithms exhibit different trends as iteration progresses. For the number of iterations and the loss value of the EKF algorithm's loss function, we can see that as the number of iterations increases, the loss value gradually decreases. This indicates that the EKF algorithm gradually optimizes the model during the iteration process, resulting in a continuous decrease in loss values. For the loss function loss value of YOLOv5 algorithm, it can also be seen that as the number of iterations increases, the loss value gradually decreases. The YOLOv5 algorithm can continuously optimize the model and reduce the loss value during the iteration process. For the loss function loss value of the algorithm obtained in this study, it can be observed that the loss value rapidly decreases during the initial iteration, and then remains within a stable range in subsequent iterations. This indicates that the algorithm obtained in this study can effectively learn and optimize the model during the initial iteration, and then basically reach a relatively optimal state in subsequent iterations.

D. ACCURACY COMPARISON AND ANALYSIS

Precision comparative analysis is one of the key indicators for evaluating the performance of an attitude recognition model, usually measured by F-score and accuracy. I compared the average accuracy of two models using and not using Kalman filtering on different test datasets. For this research topic, my analysis mainly focuses on the comparison of model accuracy. Here, I will evaluate and compare with other existing pose recognition models. We clearly evaluate the measurement indicators of accuracy. In motion scene pose recognition, accuracy refers to the ability of objects identified as real poses to be correctly recognized in prediction, and recall refers to the proportion of objects predicted as real poses to be correctly predicted. We compare the accuracy of our motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm with other comparative models based on experimental results and analysis. The experimental results are shown in Figure 9 below.

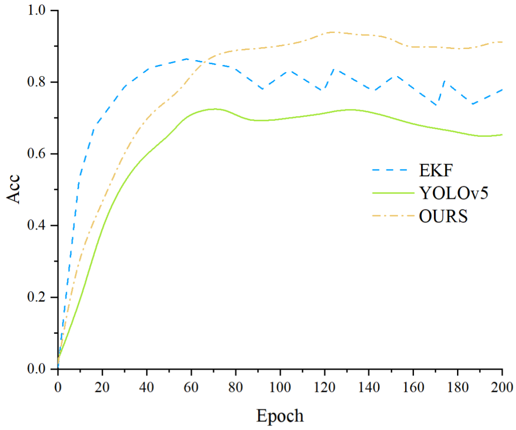

FIGURE 9. Precision comparison analysis.

Based on the data provided in Figure 9, we can observe the following situations. For the loss function accuracy of the iterative EKF algorithm, as the number of iterations increases, the accuracy gradually increases from 0.00594 to 0.83984, and then slightly decreases to 0.77504. The overall trend is gradually improving. For the loss function accuracy of the iterative YOLOv5 algorithm, as the number of iterations increases, the accuracy gradually increases from 0.02575 to 0.65337, and the overall trend is also gradually improving. For the iteration of the loss function accuracy of this research algorithm, as the number of iterations increases, the accuracy gradually increases from 0.01585 to 0.93039, and then gradually decreases to 0.91171. The overall trend is still gradually improving. The overall level of accuracy of the EKF algorithm's loss function is relatively low compared to the other two algorithms. It may be due to the limitations of the algorithm itself or its unadaptability when dealing with specific problems. The loss function accuracy of YOLOv5 algorithm has approached 0.6 or more after less than 50 iterations, indicating that the algorithm has achieved relatively good performance in the early stages of training. Moreover, through comparison, it was found that although the OURS algorithm achieved higher accuracy than the YOLOv5 algorithm in the later stage, the accuracy of the OURS algorithm was lower than that of the YOLOv5 algorithm in the initial iteration. In summary, the accuracy of the EKF algorithm improved rapidly in the first few iterations and then stabilized. The accuracy of the algorithm in this study showed a gradual improvement trend throughout the entire iteration process.

VI. CONCLUSION

This study proposes a motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm, and has achieved good performance in experiments. This article adopts the Kalman filtering algorithm to track and predict targets, effectively solving the problem of estimating the position and attitude of moving targets in complex backgrounds. The advantage of Kalman filtering is that it can dynamically predict the state of the target and modify it based on observation data, thereby improving the accuracy and stability of recognition. The study adopted the YOLOv5 algorithm for object detection and attitude estimation. The YOLOv5 algorithm can effectively improve the speed and accuracy of detection by transforming target recognition tasks into target detection tasks. In addition, this article also combines deep learning methods and uses pre trained neural network models to further improve the performance of attitude estimation.

By comparing the experimental results, we have drawn the following conclusion: the motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm has achieved significant improvements in accuracy and real-time performance. Compared with traditional methods, the model proposed in this paper has higher accuracy and faster processing speed in attitude recognition tasks. The experimental results show that the method proposed in this paper has good adaptability and reliability in different scenarios and complex backgrounds.

This study effectively solves the problems existing in traditional methods and achieves significant performance improvements by using a motion scene pose recognition model based on Kalman filtering and YOLOv5 algorithm. Future research can further optimize algorithms, expand application fields, and combine more data and technical means to improve the accuracy and practicality of attitude recognition.

REFERENCES

[1] Y. Li, Z. Li, Z. Guo, A. Siddique, Y. Liu, and K. Yu, "Infrared small target detection based on adaptive region growing algorithm with iterative threshold analysis," *IEEE Trans. Geosci. Remote Sens.*, 2024, doi: 10.1109/TGRS.2024.3376425.

[2] F. Xu, F. Xu, J. Xie, C.-M. Pun, H. Lu, and H. Gao, "Action recognition framework in traffic scene for autonomous driving system," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22301–22311, Nov. 2022.

[3] D. Meng, et al., "A data-driven intelligent planning model for UAVs routing networks in mobile Internet of Things," *Comput. Commun.*, vol. 179, 2021, pp. 231–241.

[4] M. Elshahawy, A. O. Aseeri, S. El-Sappagh, H. Soliman, M. Elmogy, and M. Abu-Elkheir, "Identification and classification of crowd activities," *Comput., Mater. Continua*, vol. 72, no. 1, pp. 815–832, 2022.

[5] P. Sun, X. Zhao, Y. Zhao, N. Jia, and D. Cao, "Intelligent optimization algorithm of 3D tracking technology in football player moving image analysis," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–13, Jul. 2022.

[6] J. Liu, I. W. Foged, and T. B. Moeslund, "Clothing insulation rate and metabolic rate estimation for individual thermal comfort assessment in real life," *Sensors*, vol. 22, no. 2, p. 619, Jan. 2022.

[7] Z. Wang, Y. Zheng, Z. Liu, and Y. Li, "A survey of video human behaviour recognition methodologies in the perspective of spatial–temporal," in *Proc. 2nd Int. Conf. Intell. Technol. Embedded Syst. (ICITES)*, Sep. 2022, pp. 138–147.

[8] Z. Wang, Y. Yang, Z. Liu, and Y. Zheng, "Deep neural networks in video human action recognition: A review," 2023, *arXiv:2305.15692*.

[9] I. Daramouskas, D. Meimetis, N. Patrinopoulou, V. Lappas, V. Kostopoulos, and V. Kapoulas, "Camera-based local and global target detection, tracking, and localization techniques for UAVs," *Machines*, vol. 11, no. 2, p. 315, Feb. 2023.

[10] X. Zhu, Q. Hua, W. Yan, Z. Guo, and K. Yu, "A vehicle-road urban sensing framework for collaborative content delivery in freeway-oriented vehicular networks," *IEEE Sensors J.*, vol. 24, no. 5, pp. 5662–5674, 2023.

[11] A. Asgharpoor Golroudbari, and M. H. Sabour, "Recent advancements in deep learning applications and methods for autonomous navigation: A comprehensive review," 2023, *arXiv:2302.11089*.

[12] G. Chen and L. Hong, "Research on environment perception system of quadruped robots based on LiDAR and vision," *Drones*, vol. 7, no. 5, p. 329, May 2023.

[13] L. Lin, H. He, Z. Xu, and D. Wu, "Realtime vehicle tracking method based on YOLOv5 + DeepSORT," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–11, Jun. 2023.

[14] X. Li and J. Wu, "Developing a more reliable framework for extracting traffic data from a UAV video," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12272–12283, Nov. 2023.

[15] J. Yang, S. Jiang, K. Chen, and L. Liu, "Yolov5-based rotating target pose grasping," in *Proc. 2nd Int. Conf. Algorithms, High Perform. Comput. Artif. Intell. (AHPCAI)*, Oct. 2022, pp. 367–373.

[16] J.-X. Li, Y. Li, R.-D. Zhan, X.-X. Zhang, H.-M. Huang, and Y.-H. Jiang, "Design and implementation of target detection and tracking system based on deep learning," in *Proc. 2nd Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2022, pp. 444–449.

[17] J. Liu, Q. Gu, D. Chen, and D. Yan, "VSLAM method based on object detection in dynamic environments," *Frontiers Neurorobotics*, vol. 16, Sep. 2022, Art. no. 990453.

[18] R. Gao, Z. Li, J. Li, B. Li, J. Zhang, and J. Liu, "Real-time SLAM based on dynamic feature point elimination in dynamic environment," *IEEE Access*, vol. 11, pp. 113952–113964, 2023.

[19] F. Zhang, T. Yang, Y. Bai, Y. Ning, Y. Li, J. Fan, and D. Li, "Online ground multitarget geolocation based on 3-D map construction using a UAV platform," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621817.

[20] X. Ren, M. Sun, X. Zhang, L. Liu, H. Zhou, and X. Ren, "An improved mask-RCNN algorithm for UAV TIR video stream target detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102660.

[21] T. Ward, S. Rashad, and H. Elgazzar, "Machine learning based pedestrian detection and tracking for autonomous vehicles," in *Proc. IEEE 13th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Mar. 2023, pp. 1294–1299.

[22] W. Pei, Z. Shi, and K. Gong, "Small target detection with remote sensing images based on an improved YOLOv5 algorithm," *Frontiers Neurorobot.*, vol. 16, Feb. 2023, Art. no. 1074862.

[23] H. Gong, L. Gong, T. Ma, Z. Sun, and L. Li, "AHY-SLAM: Toward faster and more accurate visual SLAM in dynamic scenes using homogenized feature extraction and object detection method," *Sensors*, vol. 23, no. 9, p. 4241, Apr. 2023.

[24] B. Yuan, W. Ma, and F. Wang, "High speed safe autonomous landing marker tracking of fixed wing drone based on deep learning," *IEEE Access*, vol. 10, pp. 80415–80436, 2022.

[25] G. Hao, J. Tan, Y. Lu, and G. Fu, "A method for indoor and outdoor collaborative localization and mapping based on multi-sensors unmanned platforms," in *Proc. 2nd Int. Symp. Sensor Technol. Control (ISSTC)*, Aug. 2023, pp. 108–113.

[26] B. Xing, Z. Yi, L. Zhang, and W. Wang, "Research on the mobile robot map-building algorithm based on multi-source fusion," *Appl. Sci.*, vol. 13, no. 15, p. 8932, Aug. 2023.

[27] Q. Liu, S. Wang, X. He, and Y. Liu, "Pear flower recognition based on YOLO V5S target detection model in complex orchard scenes," in *Proc. Int. Conf. Guid., Navigat. Control*. Singapore: Springer, 2022, pp. 5961–5970.

[28] S. Unar, Y. Su, P. Liu, L. Teng, Y. Wang, and X. Fu, "An intelligent system to sense textual cues for location assistance in autonomous vehicles," *Sensors*, vol. 23, no. 9, p. 4537, May 2023.

[29] Y. Zheng, C. Zheng, X. Zhang, F. Chen, Z. Chen, and S. Zhao, "Detection, localization, and tracking of multiple MAVs with panoramic stereo camera networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 2, pp. 1226–1243, Apr. 2023.

[30] L. Zhu, H. Lin, X. Chen, W. Liang, Z. Cheng, D. Shao, H. Yu, Y. Zheng, and W. Ma, "Indoor robot localization based on visual perception and on particle filter algorithm of increasing priority particles," in *Proc. 7th Int. Conf. Cyber Secur. Inf. Eng.*, Sep. 2022, pp. 1021–1026.

TINGTING WU received the B.S. degree in computer science and technology from Chongqing Jiaotong University, China, in 2007, and the M.E. degree in master of physical education and training from Southwest University, China, in 2010. She currently works with Xihua University. Her research interests include computer vision, smart sports, and sports education.

XINGFENG FAN received the bachelor's degree from Chongqing Three Gorges University, Chongqing, China, in 2007, and the B.E. degree from China University of Political Science and Law, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree. He was a General Manager with Chongqing VIP Information Company Ltd., from 2011 to 2020. Since 2021, he has been the Director of Chaxin Science and Technology Center of Liangjiang New Area, Chongqing. His research interests include big data analysis, information management, and industrial informatics.