Volleyball Motion Analysis Model Based on GCN and Cross-View 3D Posture Tracking

Hongsi Han[1], Jinming Chang[2]*
Foundation College, Shaanxi Business College, Xi'an, China[1]
Pre-normal College, Shaanxi Business College, Xi'an, China[2]

***Abstract*—The tracking of motion targets occupies a central position in sports video analysis. To further understand athletes' movements, analyze game strategies, and evaluate sports performance, a 3D posture estimation and tracking model is designed based on Graphical Convolutional Neural Network and the concept of "cross-vision". The outcomes revealed that the loss function curve of the 3D tracking model designed for the study had the fastest convergence with a minimum convergence value of 0.02. The average precision mean values for the four different publicly available datasets were above 0.90. The maximum improvement reached 21.06% and the minimum average absolute percentage error was 0.153. The higher order tracking accuracy of the model reached 0.982. Association intersection over union was 0.979. Association accuracy and detection accuracy were 0.970 and 0.965 respectively. During the volleyball video analysis, the tracking accuracy and tracking precision reached 89.53% and 90.05%, respectively, with a tracking speed of 33.42 fps. Meanwhile, the method's trajectory tracking completeness was always maintained at a high level, with its posture estimation correctness reaching 0.979. Mostly tracked and mostly lost confirmed the tracking ability of the method in a long time and cross-view state with high model robustness. This study helps to promote the development and application of related technologies, promote the intelligent development of volleyball in training, competition and analysis, and improve the efficiency of the sport and the level of competition.**

Keywords*—Graphical Convolutional Neural Network; posture estimation; volleyball; motion analysis model; 3D tracking

I. Introduction

In recent years, AI vision technology has begun to be widely used in the sports industry, accelerating the change of the sports industry. This is of key significance in enhancing the level of competitive sports, promoting national fitness activities, expanding the scale of the sports industry, and accelerating the dissemination of sports culture [1-2]. Artificial intelligence technology can develop a more personalized and precise training plan by collecting the training and competition videos of athletes and using visual technology to identify and analyze the movements. Athletes may benefit from it by improving their technical movement proficiency and training effectiveness. Simultaneously, a great deal of game data analysis helps to increase the effectiveness of tactics and strategies [3–4]. Multiple object tracking (MOT) is an important branch in the field of sports video (SV) analytics, which involves techniques such as human activity recognition (HAR), position estimation and target localization [5]. By tracking sports targets, such as players, balls, etc., key data such as their positions, velocities, accelerations, etc., can be obtained in real-time, providing accurate data support for action analysis, tactics development and athletes' performance evaluation [6]. However, there are still a range of practical challenges for MOTs in the sports arena. On the one hand, lighting changes, background interference and occlusion in real ball SVs increase the difficulty of target tracking (TT). Players between the same team are extremely similar in appearance, leading to easy identity exchange of tracking targets. At the same time, the changing body postures of the movement lead to the difficulty of existing MOT methods to distinguish and track the target only by relying on the appearance features [7-8]. On the other hand, multi-camera views in real sports arenas require precise calibration and alignment to ensure that images from different camera views can be accurately aligned and fused. Meanwhile, in order to capture key events, the multi-camera system needs to switch viewpoints frequently, which increases the difficulty of tracking targets continuously and accurately. Therefore, there is a need to develop new MOT technologies to meet the real challenges.

In order to solve the difficulty of target tracking due to the similarity of appearance and constant change of position in multi-camera viewpoints, the study takes volleyball as an example, and proposes a 3D posture estimation (3D-PE) and tracking model (TM) based on graph convolutional network (GCN) and cross-view matching for the TT difficulties caused by the athletes' similar appearance and changing postures. The study is expected to significantly improve the accuracy of athlete tracking by introducing 3D pose estimation and cross-view matching techniques, which will help to deeply analyse athletes' movement characteristics, exercise habits and potential problems, as well as promote the development and innovation of related technologies.

The research innovatively proposes a 3D pose estimation and tracking model based on GCN and cross-view matching, which provides a new technical idea and method for multi-camera viewpoint sports MOT, which can be realised for the expansion of applications in ball sports. The study provides a new theoretical perspective for the field of video analysis by deeply exploring the target tracking in multiphase viewpoints, which can help to improve the athletes' competitive level and provide a more solid technical support for the sports field.

The research is broken up into five sections. In Section II, the state of the art of motion recognition, posture estimation and visual tracking is summarized. In Section III and Section IV a two-dimensional TM for athletes with a single camera and a three-dimensional posture estimation and TM with cross-view matching are designed. In the third part, the performance

of the 3D-PE and TM is tested and analyzed. Section V summarizes the main conclusions and future work of the study.

II. Literature Review

HAR has become one of the popular researches in the field of Artificial Intelligence, HAR is a complex and multidimensional problem, which has been extensively studied by many researchers and scholars. DL techniques have application limitations in monitoring and recognizing elderly people living alone in the face of estimating missing or rare poses in the training dataset. To solve the 3D-PE fuzzy problem, Kim et al. [9] presented a loss function (LF) for the center of mass deviation from the center of the supporting foot and a penalty function for the range of rotation of the appropriate joint angle. The experimental results indicated that the average joint coordinate difference for posture estimation of this method was 0.097m with an execution time of 0.033s per frame. The wide range of human behavioral variations in daily life increases the difficulty of HAR recognition. Khan et al. [10] designed a fully automated HAR model by fusing deep neural networks (DNN), multi-view features, and a plain Bayesian classifier. The outcomes indicated that the maximum accuracy of the model is up to 99.4%. Changes in movements and different viewpoints cause difficulties in recognizing HAR. Guddeti [11] designed a multiple-learning framework for action data in depth and skeleton format from the perspective of multimodal visual data fusion. The framework could extract effective spatiotemporal features from skeleton data and utilize attention-guided DL techniques to accomplish model classification. According to experimental findings, this method's accuracy in multi-view datasets can reach 89.75%. It has been proved that GCN can be better for action recognition based on human skeleton. Aiming at the problems of GCN and the complexity between joints, Yang et al. [12] designed a hybrid network. The network integrated GCN and convolutional neural network (CNN), which could focus on the structural information and the complex relationship between nodes at the same time. The public dataset verified the superiority of the method. HAR is a key technology in wearable sensors and mobile technology, but existing HAR frameworks are only developed for a single data modality. To effectively recognize activities, Islam et al. [13] combined a multi-head CNN with a convolutional long and short-term memory network to analyze visual data and multi-source sensor information. The technique performs better in the multimodal HAR framework, according to the experimental data. HAR classification precision based on wearable sensor data is still insufficient, so Sarkar et al. [14] proposed a new hybrid HAR architecture. The architecture integrated spatial attention-assisted CNNs, filters, and techniques such as genetic algorithms and k-nearest neighbor classifiers. The public dataset confirmed the high recognition precision of the method. Due to its academic and commercial potential, MOT is a fundamental computer vision task that has drawn a lot of attention. Its associated multi-target detection, recognition, and tracking have become the focus of computer vision research and are being used more and more in a variety of fields. Li et al. [15] conducted a study on table tennis trajectory tracking in sports and designed a table tennis trajectory extraction network based on a target detection algorithm. The network incorporated a feature reuse module, enhanced the feature richness of the feature mapping using the Transformer model, and was lightweighted. The experimental results indicated that the network had a detection accuracy of 89.1% for target localization. Zhang and Dai [16] designed a model for tracking athletes' motion trajectories based on computer vision technology. Firstly, the study acquired a motion target based on a generalized background eliminator, and then used a kernelized correlation filter to construct a TT model by fusing it with relevant depth information. Through experiments on gymnasts and badminton players, it was found that the model could effectively reduce the motion trajectory prediction error. Mukhtar and Khan [17] constructed a new MOT method based on vision-Transformer architecture, variable scale pyramid, recursive pyramid structure, spatiotemporal memory encoder, and spatiotemporal memory decoder. The method could predict the object state through the attention mechanism. According to the trial results, the method's performance was much enhanced, and the ID switching rate was lowered by 21.05% and 5.79%, respectively. The traditional correlation filtering conventional feature technique was difficult to fully express the variable target morphology in complex scenes during the process of target localization, which led to inaccurate target localization. In this regard, Liu et al. [18] realized high quality visual monitoring and localization based on location fusion mechanism based on visual cognition flows (LFVC). In contrast to the most advanced visual tracking algorithms now in use for intricate scenes, this approach demonstrated superior performance at a reduced computational expense. An et al. [19] suggested a robust UAV TM based on dynamic feature weight selection to increase the flexibility of visual TT to complicated settings. The model contained multiple weights for different features and utilized dynamic feature weight selection to provide model tracking performance. The model fared better in experiments than previous cutting-edge trackers. The Kernel Correlation Filter (KCF) algorithm has achieved good results in short-term visual TT. To realize long-term TT with target occlusion or loss, Fan et al. [20] designed a long-term KCF and accelerated robust feature TT algorithm. Target matching was accomplished by the algorithm by introducing a random sample consistent target retrieval matching approach. The experimental results confirmed that the method could realize long-term stable TT. In summary, domestic and international research on HAR, positional analysis and MOT has made technical breakthroughs in terms of the main performance. Motion TT relies on the synergy of multiple technologies such as motion recognition, position estimation and target localization. However, the complex and changing game environment, the fast movement of motion targets and the occlusion of deformed other athletes or objects lead to the performance of the existing bit-pose analysis and MOT in SV still needs to be improved. In this regard, the study unfolds the design of 3D TT model based on DL and cross-view.

III. GCN-based and Cross-view 3D Attitude Tracking Technology

To solve the multi-camera multi-volleyball player tracking problem, facing the complex sports field environment and similar player identity information, the study firstly designs a

2D athlete TM based on GCN and posture alignment. Then, the 3D-PE and posture TM are proposed on this basis.

A. Two-Dimensional Tracking Modeling of Athletes with a Single Camera

Volleyball sports action analysis needs to recognize and track the player's position on the playing field and estimate the movement posture, but the similarity of player identity information and the large variation of movement posture increase the difficulty of MOT [21-22]. Therefore, the study realizes the recognition and tracking of players' movements from the perspective of appearance feature representation and contextual information differentiation, and the proposed 2D TM framework is shown in Fig. 1.

In Fig. 1, the 2D TM mainly consists of player detection and posture estimation, feature extraction, contextual graph model and similarity association matching module. The study adopts posture alignment to complete the feature extraction, including extracting the global feature map $M_g \in \mathbb{R}^{C\times H\times W}$ of the player image to construct the global feature branch, and extracting the posture heat map $M_{p_i} \in \mathbb{R}^{H\times W}$ to construct the posture alignment feature branch. Among them, $C\times H\times W$ denotes the feature map channels × height × width. K denotes the number of joint points, $i \in K$. The feature extraction model based on pose alignment is shown in Fig. 2.

In Fig. 2, the feature extraction framework uses a variant of residual network (ResNet), ResNet50, to extract $M_g \in \mathbb{R}^{C\times H\times W}$. With 50 layers deep, ResNet50 is able to resolve gradient vanishing and gradient explosion issues in DNN training. The framework uses cascaded pyramid network (CPN) to extract $M_{p_i} \in \mathbb{R}^{H\times W}$. CPN is a common DL model for posture estimation, which can effectively recognize key points in human postures, including the head, shoulders, elbows, wrists, and other key parts [23-24]. CPN utilizes cascading pyramids to construct multi-scale feature representations, which can realize human posture estimation at different scales. The CPN structure composition is shown in Fig. 3.

In Fig. 3, the CPN ontology includes GlobalNet and RefineNet, GlobalNet completes the coarse extraction of key points, and RefineNet completes the fusion of different layers of information to obtain more comprehensive and accurate posture estimation results [25]. The feature map process of the feature extraction framework is shown in Eq. (1).

$$\begin{cases} F_g = GAP\left(M_g\right) \\ F_p = GMP\left(\left\{f_{p_i}\right\}_{i=1}^{K}\right) = GMP\left(GAP\left\{M_g \otimes m_{p_i}\right\}_{i=1}^{K}\right) \end{cases} \quad (1)$$

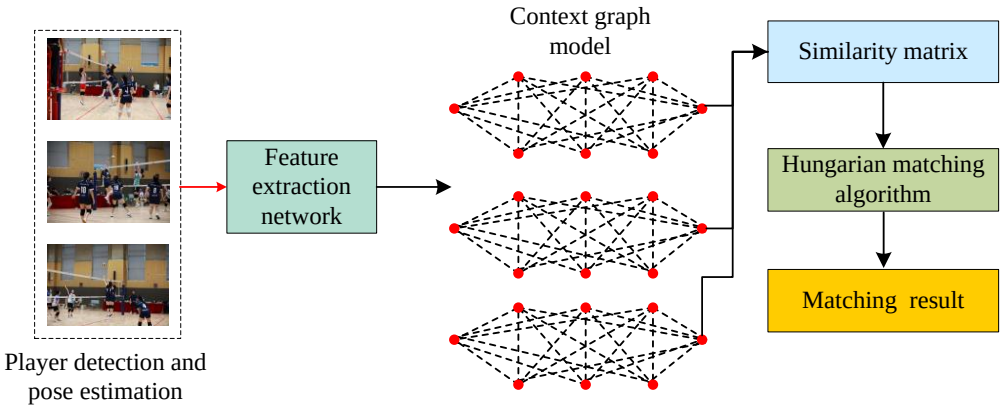


Fig. 1. Player 2D tracking model framework.

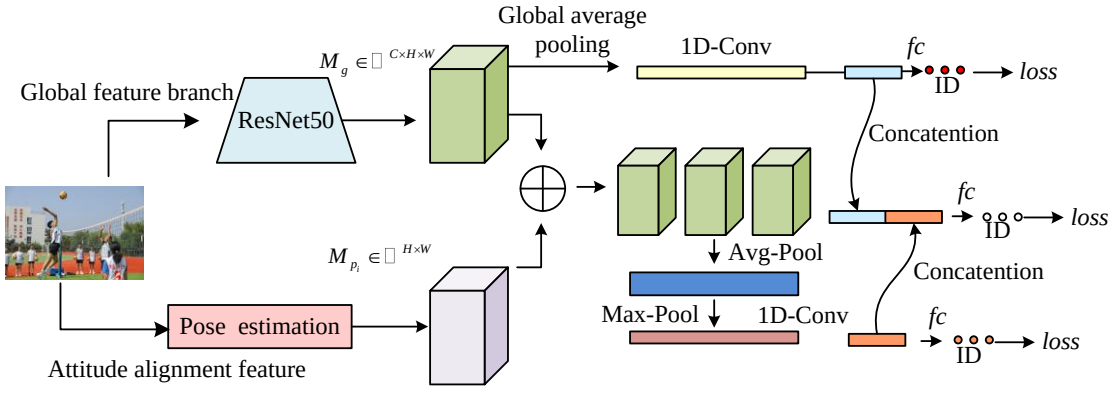


Fig. 2. Schematic diagram of the feature extraction model based on attitude alignment.

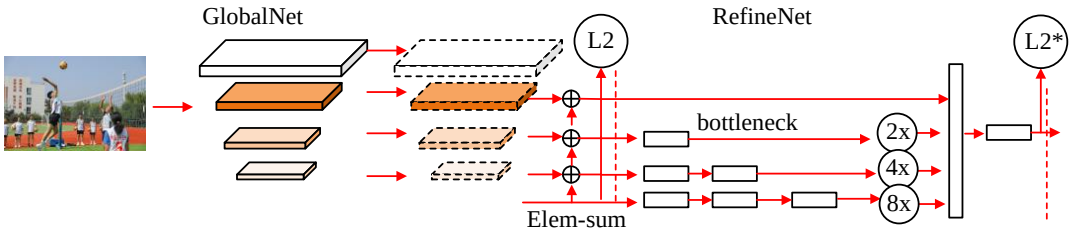


Fig. 3. Schematic diagram of CPN structure composition.

In Eq. (1), F_g and F_p denote the global features extracted from the global feature map and the pose heat map and the pose-aligned features, respectively. *GAP* and *GMP* denote the global maximum and global average pooling respectively. m_{p_i} is the pose heat map. f_{p_i} denotes the corresponding feature. Considering that there are various kinds of occlusions in the real field environment, the feature extraction framework splices the two kinds of features as the final feature extraction result. The splicing feature F_{cat} calculation process is shown in Eq. (2).

$$F_{cat} = F_g \left\| F_p \right. \tag{2}$$

The end of the feature extraction framework uses a fully connected layer to predict features. Eq. (3) computes the LF of the network.

$$loss = loss_{cat} + loss_g + loss_p \tag{3}$$

In Eq. (3), $loss_{cat}$, $loss_g$ and $loss_p$ denote the LFs corresponding to splicing features, global features and pose alignment features, respectively. The context graph model first learns the target athlete and the nearest neighbor athletes to complete the construction of the context graph $\wp = (V, E)$. V and E represent the "nodes" and "edges" respectively. Then GCN is used to learn the similarity of different athletes. A DL model called GCN is used to process graph-structured data. It carries out feature extraction and learning of node interactions and connections. GCN introduces a process similar to convolutional operation in graph data to update the representation of nodes through information transfer and aggregation between neighboring nodes to achieve learning and inference of graph-structured data [26-27]. The target athlete is regarded as the master node, the nearest neighbor athletes are regarded as branch nodes, and the branch nodes are connected to the master node. The node feature x_i expression is shown in Eq. (4).

$$x_i = \begin{cases} F_{cat} \in \square^d & i \in \{1,2,...,n\} \\ 0^d & i \in \{n+1,...,N\} \end{cases} \tag{4}$$

In Eq. (4), n denotes the number of athletes. N denotes the preset constant, $n \le N$. The adjacency matrix $A \in \square^{N \times N}$ is used to represent the "edges", as shown in Eq. (5).

$$A_{i,j} = \begin{cases} 1, & i = 1 \text{ or } i = j \\ 0, & otherwise \end{cases} \quad (i, j \in \{1,2,...,N\}) \tag{5}$$

Fig. 4 depicts the node feature update mechanism. The study uses GCN to integrate node information. The GCN working mechanism mainly contains two parts: information dissemination and information aggregation. Node x_i is coded as $x_i^{(l)}$ in the propagation process and its coded as x_i in the information aggregation process. The node features obtained after GCN integration have a stronger ability to represent contextual features.

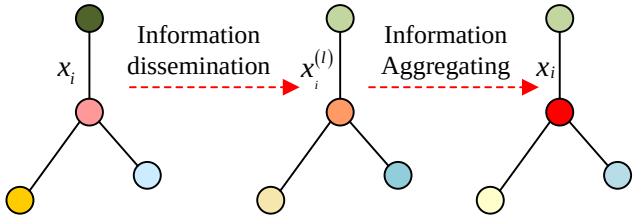


Fig. 4. Feature update mechanism for GCN integrating node information.

There is a difference in the importance of different nearest neighbor athletes to the target athlete. The study introduces the linear distance between athletes as weights into the contextual graph model. The feature aggregation process is shown in Eq. (6).

$$X^{(l+1)} \in \sigma\left(D^{-1} A X^{(l)} W^{(l)}\right) \tag{6}$$

In Eq. (6), D and A denote the distance matrix and the normalization result of "edge" A, respectively. $X^{(l)}$ denotes the set of node features $x_i^{(l)}$. $W^{(l)}$ denotes the network parameter matrix, and σ denotes the nonlinear activation function. Therefore, the GCN network with the introduction of weighting information has a stronger feature representation ability. The study uses cosine distance to measure the similarity between athletes. The cosine LF is shown in Eq. (7).

$$loss(x_1, x_2, y) = \begin{cases} 1 - \cos(x_1, x_2) & y=1 \\ \max(0, \cos(x_1, x_2) - margin) & y=-1 \end{cases} \tag{7}$$

In Eq. (7), y denotes true similarity labeling and $y = 1$ denotes similarity. x_1, x_2 denotes different athletes. The similarity association matching module mainly accomplishes the matching between the current detection and the tracking trajectory. The working mechanism is shown in Fig. 5.

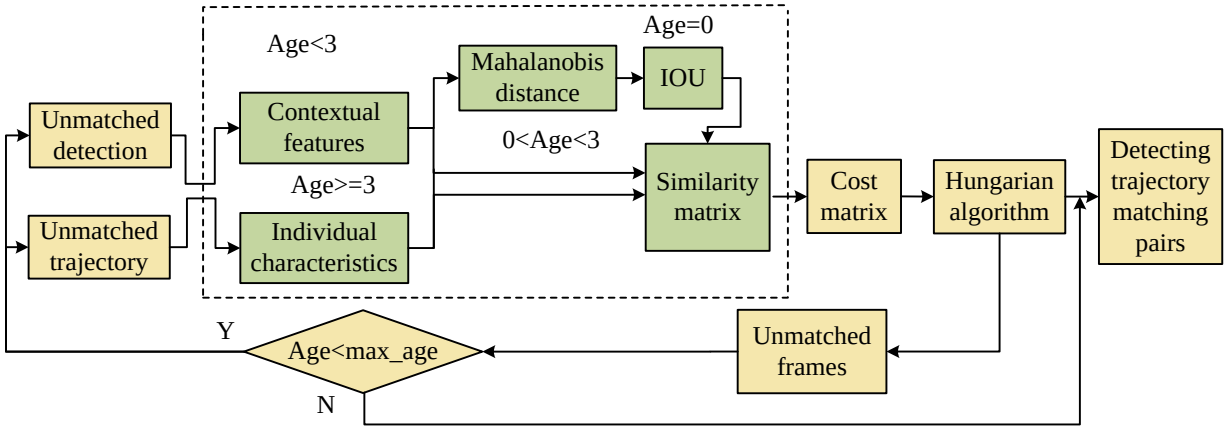


Fig. 5. Working mechanism of similarity correlation matching module.

In F. 5, "age" indicates the number of unmatched frames in the trajectory. The matching mechanism involves four different kinds of matching information, including contextual features, individual features, martensitic distance of motion states, and intersection ratio. Finally, Hungarian algorithm (HA) is used to complete the matching solution [28-29]. HA is mainly used for the solution of assignment problems and is applicable to problems related to bipartite graphs.

B. 3D-PE and Tracking Model Construction Based on Cross-View Matching

Real field videos are mainly captured by multiple cameras. To solve the complexity and challenge of real field video due to the multi-camera shooting environment, the research unfolds the 3D TM design of the target in multi-phase view based on single-camera TT. The model framework is shown in Fig. 6.

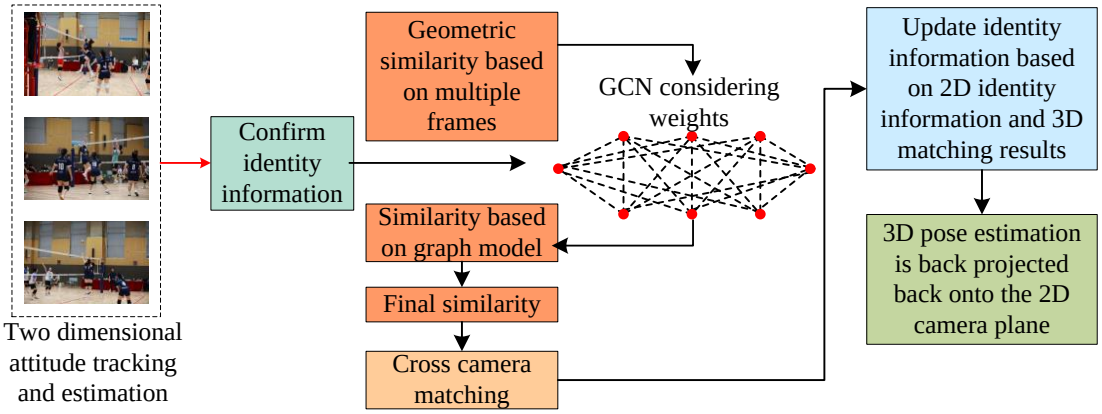


Fig. 6. Cross view matching 3D pose estimation and tracking model architecture.

In Fig. 6, the architecture contains the main modules of 2D tracking, 2D posture estimation, cross-camera matching, 3D tracking and posture estimation. Among them, 2D tracking is accomplished by the architecture shown in Fig. 1, and 2D posture estimation is accomplished by CPN network. The core of cross-camera matching lies in the similarity measure and matching of athletes from different cameras. Therefore, the study considers both geometric constraints and appearance features for similarity metrics.

Define that there are C cameras and M athletes under cross-view matching. The similarity scores between the athletes form matrix $A \in \mathbb{R}^{M \times M}$, and output the matching result $P \in \{0,1\}^{M \times M}$. "1" indicates the same athlete, and a "0" indicates different athletes. The study introduces the principle of "polar line constraint", which judges the similarity of athletes by determining whether the corresponding joints of different athletes are "homonymous image points". The geometric distance $D_g(i,j)$ across the camera of different athletes is calculated in Eq. (8).

$$D_g(i,j) = \frac{1}{2Q}\sum_{q=1}^{Q}\left(d_g\left(p_i^q, l_{ij}\left(p_j^q\right)\right) + d_g\left(p_j^q, l_{ji}\left(p_i^q\right)\right)\right) \quad (8)$$

In Eq. (8), p_i and p_j denote the two-dimensional attitude coordinates of different athletes. Q denotes the number of joint points, $q \in Q$. $l_{ij}\left(p_j^q\right)$ denotes the pole line corresponding to point p_j^q and point p_i^q in different cameras. $d_g(\cdot, l)$ denotes the distance from the point to the polar line. The distance matrix $D_g \in \mathbb{R}^{M \times M}$ is calculated from Eq. (8). The sigmoid function is used to normalize the distance matrix, and the process is shown in Eq. (9).

$$D_g(i,j)=\begin{cases}\tau_g & D_g(i,j)\geq\tau_g\\ D_g(i,j) & other\end{cases} \tag{9}$$

In Eq. (9), $D_g(i,j)$ denotes the normalized distance matrix. τ_g denotes the set cross-camera distance threshold. Standard deviation normalization is also required due to the presence of some anomalous values taken. The calculation process is shown in Eq. (10).

$$D_g(i,j)=\left(D_g-\mu\right)/\sigma \tag{10}$$

In Eq. (10), μ, σ denote the mean and standard deviation, respectively. The normalized similarity matrix A_g is shown in Eq. (11).

$$A_g=sigmoid\left(-D_g\right) \tag{11}$$

Relying on a single similarity matrix cannot distinguish the athlete identity information well. The study introduces two-dimensional trajectory information on this basis and proposes a similarity matrix based on trajectory information. The calculation process is shown in Eq. (12).

$$A_g^T=sigmoid\left(-C_g^T\right) \tag{12}$$

In Eq. (12), T denotes the past video frame. C_g^T denotes the count matrix. The study used GCN to construct a graph model across viewpoints for matching between athletes. Fig. 7 illustrates the procedure for determining how similar athletes are from various angles using the graph model.

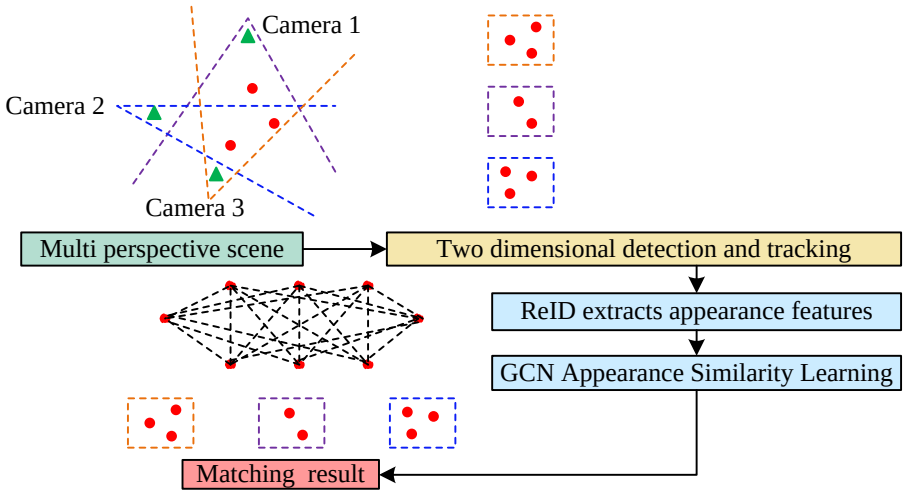


Fig. 7. Figure model learning cross perspective athlete similarity.

In Fig. 7, the study first extracts the appearance features of athletes using pedestrian re-identification (ReID) A_a. The obtained appearance feature vectors form the nodes of the graph model, and the athletes with different camera planes are connected to form the edges of the graph model. The process of applying the graph model to determine how similar athletes are from different perspectives is shown in Fig. 7. The realization process is shown in Eq. (13).

$$\begin{aligned}H^{(l+1)}&=\sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)=\\ &\sigma\left(D^{-\frac{1}{2}}\left(A_g^T\cdot A+I\right)D^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)\end{aligned} \tag{13}$$

In Eq. (13), D denotes the degree matrix. A denotes the self-looping adjacency matrix. $H^{(l+1)}$ and $H^{(l)}$ denote the node features of layer $l+1$ and layer l, respectively. After graph convolution, the more robust appearance similarity matrix A_a^K is obtained, and finally the matching is completed based on the average value of the two similarity matrices A_{g+a}. The calculation process is shown in Eq. (14).

$$A_{g+a}=\left(A_g^T+A_a^K\right)/2 \tag{14}$$

According to the study, matching athletes across viewpoint states is an optimization problem that has to be resolved. The matching process is shown in Eq. (15).

$$\min_P f(P)=-\sum_{i=1}^{M}\sum_{j=1}^{M}\left(A_{ij}\cdot P_{ij}\right)+\lambda rank(P) \tag{15}$$

In Eq. (15), A, P denote the similarity matrix, binary matching matrix respectively. $rank(P)$ denotes the rank of matrix P. To ensure the cyclic consistency of the matching results, $rank(P)$ needs to satisfy certain constraints. That is, $rank(P)\leq m$ and m denote the number of athletes in a multi-camera scene. After completing the 2D tracking and cross-camera matching, the study utilizes the triangulation algorithm to further obtain the 3D pose information. The triangulation algorithm recovers the coordinates of the 3D points from the projected coordinates of the two cameras and the transformation matrix [30-31]. The computational procedure for projection to the normalization plane is shown in Eq. (16).

$$d\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = TX \tag{16}$$

In Eq. (16), d is the depth of the 3D point X. T is the transformation matrix of the camera. Finally, the results of the 2D processing are combined to realize the athlete tracking in 3D cross-camera view. The new tracking IDs are clustered for different athletes.

IV. 3D-PE AND TRACKING MODEL PERFORMANCE ANALYSIS

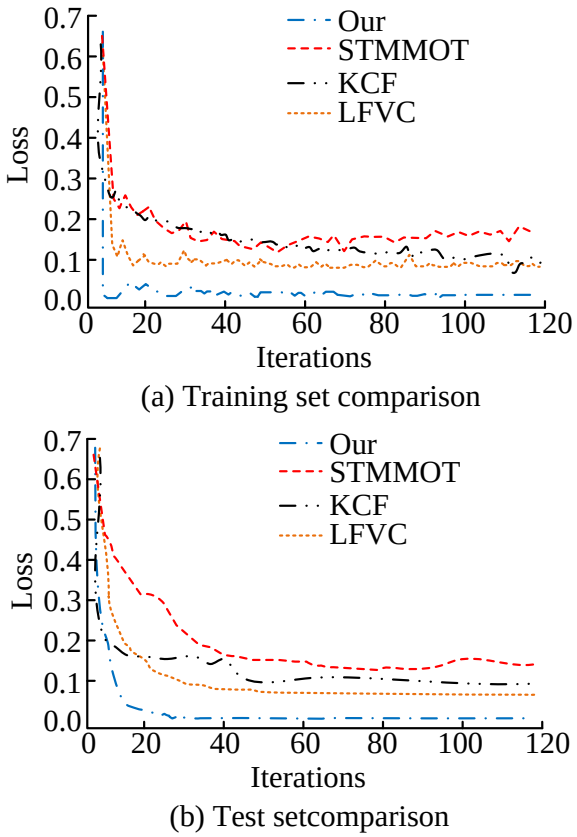
To verify the validity of the 3D-PE and TM, the study is centered on the testing of the model performance and application effects, and the results are analyzed and discussed.

A. 3D-PE and Tracking Model Performance Testing

The study launches performance testing and application analysis experiments, which are conducted based on Centos 7 operating system. The DL framework is Tensorflow-gpu-1.10.3. The graphics card is Tesla P100-PCIE with 125 GB of RAM. The central processing unit is 2.7 GHz dual core Intel Core i5. SportsMOT, UA-DETRAC dataset, DukeMTMC and MOTChallenge20 are selected as experimental datasets. SportsMOT is a large-scale multi-TT dataset for multi-sport scenarios, consisting of 240 videos. In total, it consists of about 150,000 frames and 1.6 million labeled borders. These are tracking targets that are fast changing, variable and have a similar but distinguishable appearance. The UA-DETRAC dataset is a collection of video clips containing 10 hours of video footage from 24 different roadways taken with a Cannon EOS 550D camera. The videos are shot at 25 frames per second with a resolution of 960 x 540 pixels, while three different levels of occlusion are included in the dataset. The DukeMTMC dataset is a pedestrian recognition dataset that provides a collection of more than 7,000 single-camera tracks recorded by eight synchronized cameras and more than 2,700 individual characters. MOTChallenge20 focuses on the pedestrian multi-TT task. The data required for selecting the experiments are divided into training and test sets in the ratio of 8:2. The comparative analysis models include depth-informed KCF tracking method based on literature [16], spatiotemporal memory networks and multi-scale attention pyramids (STMMOT) from literature [17] and LFVC from literature [18]. The LF curves of different TMs in the test set and training set are shown in Fig. 8.

The LF curves are a useful tool for evaluating the model's generalization capacity and learning efficiency since they may show how the model's loss value changes during training. The LF curve of the improved GCN integration model designed by the study converges with the smallest number of iterations and the smallest convergence value. On the training set, the research-designed model reduces 0.27 compared to the STMMOT model, 0.08 compared to the LFVC model, and 0.07 compared to the KCF model. Moreover, the fluctuation of the LF curves of the other models is more obvious, and the research-designed model has a more stable loss value. Fig. 9 displays the mean absolute percentage error (MAPE) and mean average precision (MAP) for each of the models.

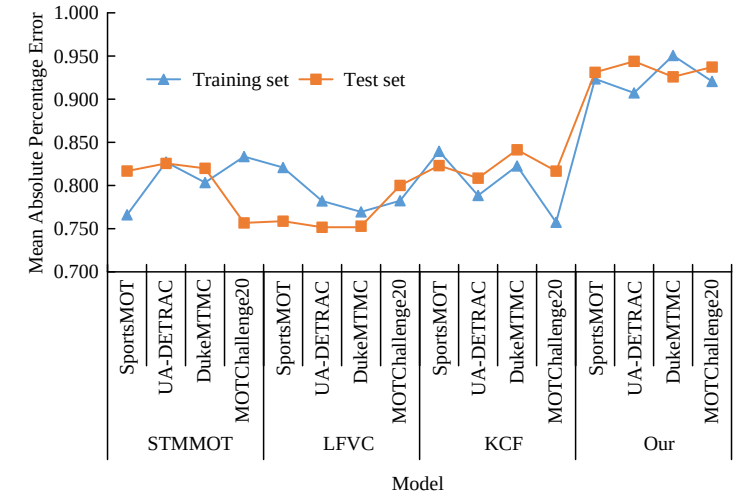
In Fig. 9(a), the optimal performance of the research-designed method is achieved on four different datasets. The value levels are all above 0.90, with the best model performance. Its maximum improvement can be up to 21.06%, 16.28%, and 17.20% compared to STMMOT model, LFVC model, and KCF model, respectively. From Fig. 9(b), the method designed by the study is at the lowest level in terms of error take, with a minimum MAPE of only 0.153 on the SportsMOT dataset, whereas the other three models take MAPE values above the 0.250 level. The target localization precision of the model is higher and the discrepancy between the tracking results and the true value is smaller when the model's MAPE value is lower. The experiments examined the higher order tracking accuracy (HOTA) of different models and the corresponding association intersection over union (AssIOU), association accuracy (AssA), and detection accuracy (DetA). Table I displays the experiment's statistical findings.



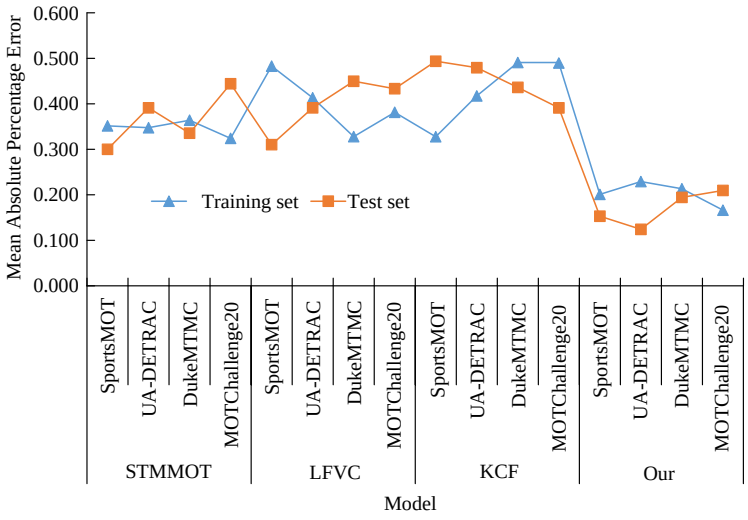
(a) Training set comparison

(b) Test setcomparison

Fig. 8. Comparison of LF curves for different models.



(a) MAP



(b) MAPE

Fig. 9. Comparison of average precision mean and average absolute percentage error for different models.

TABLE I. COMPARISON OF DETECTION AND ASSOCIATION PERFORMANCE OF DIFFERENT MODELS

Model	Index	SportsMOT	UA-DETRAC	DukeMTMC	MOTChallenge20
Our	AssIOU	0.929	0.979	0.970	0.934
	AssA	0.941	0.970	0.964	0.937
	DetA	0.922	0.957	0.965	0.946
	HOTA	0.924	0.943	0.929	0.982
STMMOT	AssIOU	0.800	0.722	0.802	0.778
	AssA	0.754	0.797	0.728	0.802
	DetA	0.754	0.770	0.780	0.741
	HOTA	0.819	0.751	0.700	0.783
LFVC	AssIOU	0.736	0.751	0.792	0.734
	AssA	0.708	0.738	0.718	0.807
	DetA	0.759	0.712	0.819	0.804
	HOTA	0.789	0.808	0.800	0.720
KCF	AssIOU	0.783	0.735	0.726	0.747
	AssA	0.775	0.738	0.791	0.775
	DetA	0.744	0.750	0.805	0.796
	HOTA	0.823	0.770	0.794	0.779

The maximum AssIOU value of the models designed for the study is 0.979, the maximum AssIOU value of the STMMOT model is 0.802, the maximum AssIOU value of the LFVC model is 0.792, and the maximum AssIOU value of the KCF model is 0.783. AssIOU can be used to measure the degree of overlap between the predicted and real frames. Its association between the trajectory and the true trajectory in the tracking task can be reflected in the MOT task. The research design model has the highest overlap between prediction and truth. The tracking algorithm's accuracy in identifying the target is evaluated by AssA; the greater the value, the more accurate the association. The more accurately a target is recognized, the greater the DetA value, which indicates the accuracy of the connection. In the same experimental environment, the model designed by the study has the highest AssA and DetA values of 0.970 and 0.965, respectively. In addition, the model has the highest HOTA value of 0.982, which is a maximum improvement of 40.28% compared to the STMMOT model. HOTA is the accuracy of both detection and association, and the higher value taken indicates better overall tracking performance. It can be concluded that the model designed by the study has some performance advantages over the most current models.

B. 3D-PE and Tracking Model Application Analysis

A volleyball game dataset is collected at a stadium in China, synchronized by six cameras evenly distributed around the venue. The acquisition frame rate of each camera is 20fps, the resolution is set to 900*900, and the acquisition time is five minutes. The acquired data is divided into five video sequences. Comparisons of multiple object tracking accuracy (MOTA), tracking speed (fps) and multiple object tracking precision (MOTP) for different models are shown in Fig. 10.

In Fig. 10(a), the method designed by the study takes the highest level of 89.53% on MOTA. The model performance is optimal when considering the tracking accuracy, false detection and missed detection. In Fig. 10(b), the tracking speed of the model designed by the study, STMMOT model, LFVC, and KCF model are 33.42fps, 84.61fps, 102.88fps, and 153.37fps, respectively. The tracking speed is improved by up to 129.95fps, and the designed by the study has a significant advantage in the rate of processing video frames. In Fig. 10(c), the method designed by the study also has a significant advantage in MOTP with a maximum fetch of 90.05%. It has a minimum improvement of 10.06 percentage points compared to other models, with a high precision in estimating the target location. Percentage of correct keypoint (PCK) evaluates the accuracy of the model's prediction of keypoint locations in posture estimation. The higher the value of PCK the more accurate the posture estimation. Also compare the ability of the model in maintaining trajectory continuity and integrity, the experimental results are shown in Fig. 11.

In Fig. 11(a), the accuracy of key point locations predicted by the research designed method is higher compared to other models. The PCK floats roughly in the 0.85-0.95 range. In Fig. 11(b), the research-designed method summarizes the TT process. The trajectory tracking completeness is consistently high, with the vast majority of completenesses above 0.7. To further analyze the degree of contribution of the improvement strategy to the quality of human posture estimation and tracking, the study introduces the posture estimation correctness, the number of times the tracking trajectory changes its matched real identities (IDs), and the ID F1 score (ID F1). Table II displays the outcomes of the experiment.

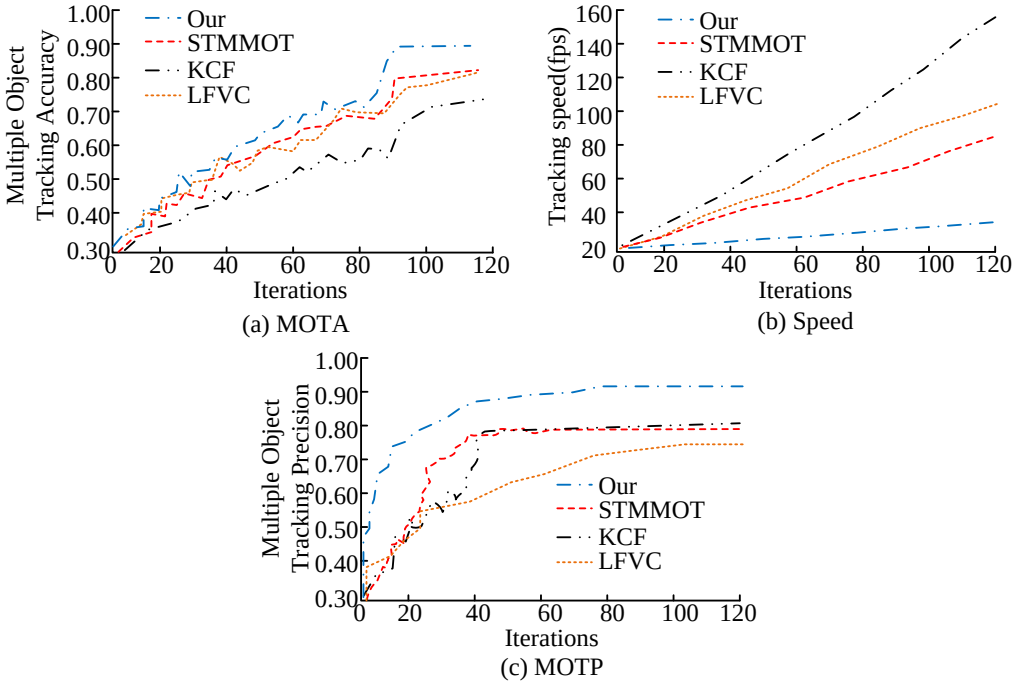


Fig. 10. Comparison of MOTA, tracking speed, and MOTP.

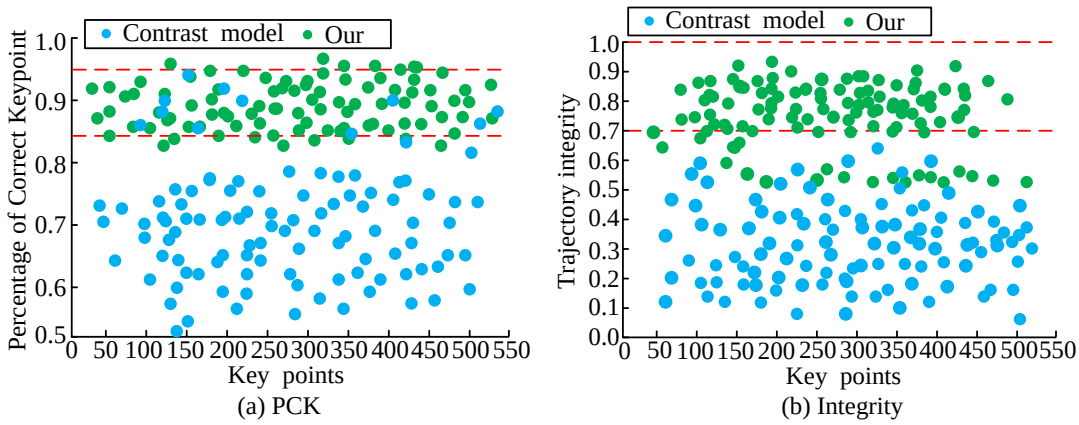


Fig. 11. Comparison between PCK and trajectory integrity preservation.

TABLE II. COMPARISON OF ATTITUDE ESTIMATION AND 3D TRACKING RESULTS

Similarity model		3D			2D		
		Accuracy	IDs	ID F1	Accuracy	IDs	ID F1
Geometric similarity	A_g	0.661	54	0.446	0.643	82	0.619
	A_g^T	0.794	19	0.713	0.697	42	0.646
Appearance similarity	A_a	0.613	135	0.606	0.706	234	0.708
	A_a^K	0.764	21	0.761	0.791	19	0.747
Geometry+Appearance	A_{g+a}	0.813	10	0.847	0.846	8	0.814
True similarity		0.979	0	0.974	0.972	0	0.981

The similarity matrix A_g^T based on the trajectory information shows a significant improvement on the estimation correctness rate, IDs, and ID F1 over the pre-improvement period. In 3D state, the correct rate is improved by 0.133, IDs are reduced by 35, and ID F1 scores are improved by 0.267. It can be observed that the improved model is more accurate in estimating the positional pose, and it is less likely to happen that the trajectory of one target is incorrectly assigned to another target, and the target identity consistency is maintained better. The introduction of GCN improved appearance similarity A_a^K likewise improves the estimation correctness, IDs, and the value of ID F1. Finally, a posture estimation correctness of 0.979 is achieved in the 3D tracking state. The IDs and ID F1 scores take the values of 0 and 0.974, respectively. The tracking algorithm designed in the study matches the real trajectories of all the targets almost perfectly. Continuing to compare the mostly tracked (MT) and mostly lost (ML) of the different models. Fig. 12 presents the outcomes of the experiment.

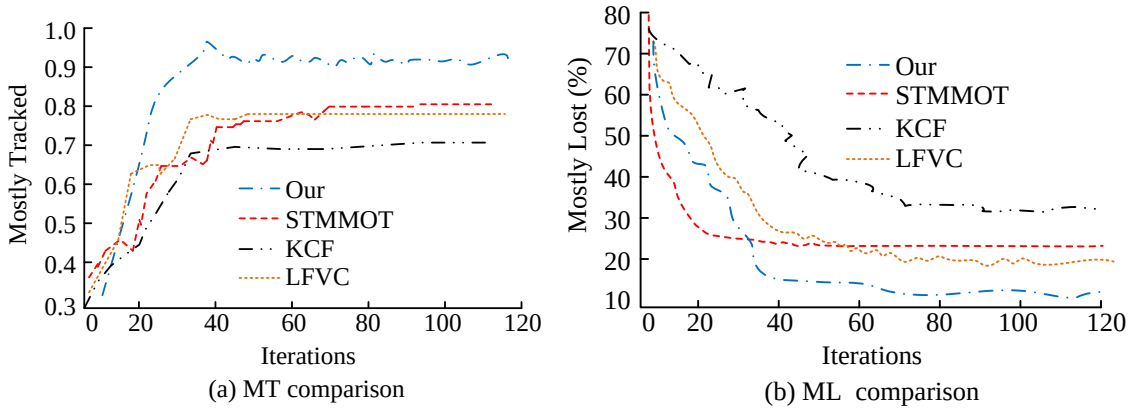
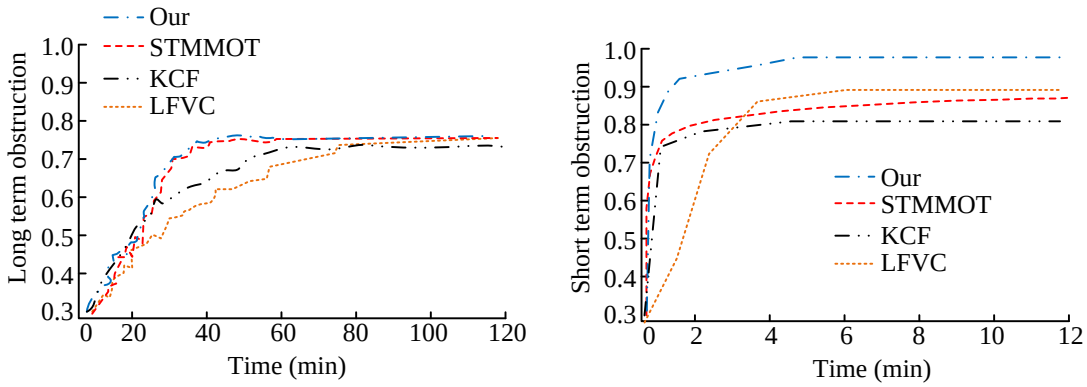


Fig. 12. Comparison of MT and ML for different models.



(a) Long term comparison (b) Short term obstruction

Fig. 13. Comparison of model robustness.

In Fig. 12(a), the research design model is able to track the target continuously, with the percentage of tracked frames exceeding 80% of the total number of video frames reaching 0.924. In contrast, the other models have MT values below the 0.90 level, with the highest value of 0.813. The research design has a better ability to maintain continuity of the TT over a long period of time. In Figure 12(b), the model of the studied design corresponds to the smallest ML, and the percentage of lost tracking exceeding 20% of the total number of video frames is only 10.36%, which maintains a strong TT stability. In summary, the two complementary metrics, MT and ML, confirm the tracking ability of the research design model over a long period of time and across viewpoint states. Finally, the robustness of the model is compared and analyzed, and the ratio of recovered tracks from short-term occlusion and long-term occlusion is compared, and the experimental results are shown in Fig. 13.

The research-designed method is still able to maintain good trajectory continuity when facing challenges such as target occlusion. The ratio of recovered tracking trajectories in short-term occlusion is above 0.90. Compared to other methods, the research design has a better performance in terms of tracking robustness. However, the ratio of recovering tracking trajectory under long-term occlusion is only 0.765. Although the performance is comparable to other models, the model still has more room for improvement.

V. Conclusion

The continuous progress of vision technology and the expansion of application scenarios have brought great convenience to human life. To cope with the complexity and richness of SVs and to improve the precision and speed of MOT, the research has modeled and analyzed the 3D position estimation and TT, and proposed a new cross-view 3D-PE and TM. The outcomes revealed that the LF values of the research design model were reduced by 0.27 compared to the STMMOT model, 0.08 compared to the LFVC model, and 0.07 compared to the KCF model. The MAP and MAPE took better values than the other baseline models. The maximum AssIOU value was 0.979. The maximum on AssA and DetA values were 0.970 and 0.965 respectively. The maximum HOTA value was 0.982. In real volleyball video analysis, the method fetched the highest level of 89.53% on MOTA. Its tracking speed was improved by up to 129.95fps, and the maximum fetch level on MOTP reached 90.05%. The improved strategy designed by the study improved the correct rate in 3D by 0.133, IDs by 35, and ID F1 scores by 0.267. The MT value achieved was 0.924, the ML minimum was only 10.36%, and the rate of short-term occlusion to recover the tracking trajectory was above 0.90. The study realizes the tracking of complex SVs, which helps to deepen the theory and methodology of cross-view 3D pose tracking techniques. However, the rate at which the model designed by the study can successfully recover the tracking trajectory under long-term occlusion is still insufficient. The algorithm still needs to be investigated in maintaining a long-time memory of the target identity.

In addition, facing real application scenarios, the study still has some limitations. First, the computational efficiency and model response speed of the model may be limited when dealing with complex scenes and large-scale data. Second, in order to improve the generalisation ability of the model, a large amount of video data from different motion scenes is required, but there are difficulties in obtaining diverse annotated datasets. Whether the model can be adapted to other motion fields needs to be further verified.

Acknowledgment

The research is supported by Special Project of Educational and Teaching Reform Research in 2024 of Shaanxi Business College: Special Project of the "Simultaneous Promotion of Five Educations" Education Mode. (27) A Research on the Path of Two-way Integration of Higher Vocational Physical Education and Curriculum Ideological and Political Education under the Background of "Simultaneous Promotion of Five Educations".

References

[1] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: A survey," Neural Comput. Appl., vol. 35, no. 5, pp. 4145-4182, October 2023.

[2] H. C. Nguyen, T. H. Nguyen, R. Scherer, and V. H. Le, "Deep learning for human activity recognition on 3d human skeleton: Survey and comparative study," Sens., vol. 23, no. 11, pp. 5121-5146, May 2023.

[3] Y. Sun, Y. Weng, B. Luo, G. Li, B. Tao, D. Jiang, and D. Chen, "Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images," IET Image Process., vol. 17, no. 4, pp. 1280-1290, December 2023.

[4] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: A review," Complex Intell. Syst., vol. 10, no. 1, pp. 1581-1606, July 2023.

[5] H. B. Mahajan, N. Uke, P. Pise, M. Shahade, V. G. Dixit, S. Bhavsar, and S. D. Deshpande, "Automatic robot Manoeuvres detection using computer vision and deep learning techniques: a perspective of internet of robotics things (IoRT)," Multimed. Tools Appl., vol. 82, no. 15, pp. 23251-23276, June 2023.

[6] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "IMoS: Intent-driven full-body motion synthesis for human-object interactions," Comput. Graph. Forum., vol. 42, no. 2, pp. 1-12, May 2023.

[7] N. Dua, S. N. Singh, V. B. Semwal, and S. K. Challa, "Inception inspired CNN-GRU hybrid network for human activity recognition," Multimed. Tools Appl., vol. 82, no. 4, pp. 5369-5403, February 2023.

[8] L. Cai, N. E. McGuire, R. Hanlon, T. A. Mooney, and Y. Girdhar, "Semi-supervised visual tracking of marine animals using autonomous underwater vehicles," Int. J. Comput. Vision, vol. 131, no. 6, pp. 1406-1427, March 2024.

[9] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, "Human pose estimation using mediapipe pose and optimization method based on a humanoid mode," Appl. Sci., vol. 13, no. 4, pp. 2700-2720, February 2023.

[10] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," Multimed. Tools Appl., vol. 83, no. 5, pp. 14885-14911, March 2024.

[11] R. M. R. Guddeti. "Human action recognition using multi-stream attention-based deep networks with heterogeneous data from overlapping sub-actions," NEURAL COMPUT APPL, vol. 36, no. 18, pp. 10681-10697, March, 2024, DOI: 10.1007/s00521-024-09630-0.

[12] W. Yang, J. Zhang, J. Cai, and Z. Xu, " HybridNet: Integrating GCN and CNN for skeleton-based action recognition," Appl. Intell., vol. 53, no. 1, pp. 574-585, April 2023.

[13] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things," Inform. Fusion, vol. 94, no. 6, pp. 17-31, June 2023.

[14] A. Sarkar, S. S. Hossain, and R. Sarkar, "Human activity recognition from sensor data using spatial attention-aided CNN with genetic algorithm," Neural Comput. Appl., vol. 35, no. 7, pp. 5165-5191, October 2023.

[15] W. Li, X. Liu, K. An, C. Qin, and Y. Cheng, "Table tennis track detection based on temporal feature multiplexing network," Sens., vol. 23, no. 3, pp. 1726-1754, February 2023.

[16] L. Zhang and H. Dai, "Motion trajectory tracking of athletes with improved depth information-based KCF tracking method," Multimed. Tools Appl., vol. 82, no. 17, pp. 26481-26493, July 2023.

[17] H. Mukhtar and M. U. G. Khan, " STMMOT: Advancing multi-object tracking through spatiotemporal memory networks and multi-scale attention pyramids," Eural Netw., vol. 168, no. 11, pp. 63-379, November 2023.

[18] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista, and J. Del Ser, "Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows," Inform. Fusion, vol. 96, no. 8, pp. 281-296, August 2023.

[19] Z. An, X. Wang, B. Li, Z. Xiang, and B. Zhang, "Robust visual tracking for UAVs with dynamic feature weight selection," Appl. Intell., vol. 53, no. 4, pp. 3836-3849, February 2023.

[20] J. Fan, X. Yang, R. Lu, W. Li, and Y. Huang, "Long-term visual tracking algorithm for UAVs based on kernel correlation filtering and SURF features," Visual Comput., vol. 39, no. 1, pp. 319-333, January 2023.

[21] D. Meimetis, I. Daramouskas, I. Perikos, and I. Hatzilygeroudis, " Real-time multiple object tracking using deep learning methods," Neural Comput. Appl., vol. 35, no. 1, pp. 89-118, January 2023.

[22] S. Ankalaki and M. N. Thippeswamy, " A novel optimized parametric hyperbolic tangent swish activation function for 1D-CNN: application of sensor-based human activity recognition and anomaly detection," Multimed. Tools Appl., vol. 83, no. 22, pp. 61789-61819, July 2024.

[23] I. Priyadarshini, R. Sharma, D. Bhatt, and M. Al-Numay, "Human activity recognition in cyber-physical systems using optimized machine learning techniques," Cluster Comput., vol. 26, no. 4, pp. 2199-2215, August 2023.

[24] T. Kamnardsiri, S. Boripuntakul, and C. Kaiket, " Computer vision-based instantaneous speed tracking system for measuring the subtask speed in the 100-meter sprinter: Development and concurrent validity study," Heliyon, vol. 10, no. 2, pp. 24086-24106, September 2024.

[25] J. Lukavský and H. S. Meyerhoff, "Gaze coherence reveals distinct tracking strategies in multiple object and multiple identity tracking," Psychon. Bull. Rev., vol. 31, no. 3, pp. 1280-1289, October 2024.

[26] A. M. Mansourian, V. Somers, C. De Vleeschouwer, and S. Kasaei, "Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking," Proc. 6th Int. Works. Multimed. Content Anal. Sport., vol. 96, no. 8, pp. 103-112, October 2023.

[27] K. Bhosle and V. Musande, "Evaluation of deep learning CNN model for recognition of devanagari digit," Artif. Intell. Appl., vol. 1, no. 2, pp. 114-118, February 2023.

[28] S. Pastel, J. Marlok, N. Bandow, and K. Witte, "Application of eye-tracking systems integrated into immersive virtual reality and possible transfer to the sports sector-A systematic review" Multimed. Tools Appl., vol. 82, no. 3, pp. 4181-4208, January 2023.

[29] Q. Gou and S. Li, "Study on the correlation between basketball players' multiple-object tracking ability and sports decision-making," PLOS ONE, vol. 18, no. 4, pp. 283965-283974, April 2023.

[30] J. Zhang, H. Huang, X. Jin, L. D. Kuang, and J. Zhang, " Siamese visual tracking based on criss-cross attention and improved head network," Multimed. Tools Appl., vol. 83, no. 1, pp. 1589-1615, January 2024.

[31] C. Wei, Y. Xiong, Q. Chen, and D. Xu, "On adaptive attitude tracking control of spacecraft: A reinforcement learning based gain tuning way with guaranteed performance," Adv. Space Res., vol. 71, no. 11, pp. 4534-4548, June 2023.