ICIC Express Letters
Part B: Applications
Volume **14**, Number **8**, August **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **803–812**

ALGORITHM RESEARCH OF HUMAN RECOGNITION AND THEIR APPLICATION TO CONJECTURE

Quanshu Song[1], Kotaro Hashikura[2], Md Abdus Samad Kamal[2]
and Kou Yamada[2]

[1]Graduate School of Science and Technology
[2]Division of Mechanical Science and Technology
Gunma University
1-5-1 Tenjincho, Kiryu 376-8515, Japan
{ t211b604; k-hashikura; maskamal; yamada }@gunma-u.ac.jp

Received December 2022; accepted February 2023

Abstract. *In this paper, we propose a specific algorithm for multi-target recognition in simple scenes using human recognition technology. In addition, we apply these methods to the field of assisting referees in football matches and prepare characteristics of their method. At present, Video Assistant Referee (VAR) technology is widely used in the world to assist judgments, but there are problems such as high price, dependence on high-end precision equipment, and little impact on referee decision-making. It is necessary to develop a simple and general-purpose auxiliary referee system for use in simple scenarios such as campuses. This requires an algorithm that can judge foul actions. This paper provides a new idea for the promotion of VAR. If it can be promoted, it is expected to shine at smaller events.*

Keywords: Machine vision, Human recognition, Alphapose, Video assistant referee, RMPE, Object recognition, Foul behavior judgment

1. **Introduction.** With the development of machine vision and artificial intelligence, the application of machine vision and artificial intelligence in life has become more and more extensive. A wonderful contest is inseparable from the efforts of the players, and it is also inseparable from the fair judgment of the referee [1]. Before the introduction of Video Assistant Referee (VAR) technology, the penalty was mainly judged by a head referee and two deputy referees through eyes to determine whether or not a player violated the rules. After the introduction of VAR technology, some more controversial scenes can be handled by playing back the pictures, which greatly reduces the possibility of misjudgment [2]. According to data provided by International Federation of Association Football (FIFA), VAR can allow referees to change decisions every 3.29 games. During the 2018 World Cup, referees' misjudgment green dropped from 5% to 0.68%. VAR also has some problems; for example, the cost is too high. The cost of the use of VAR for a season was estimated at US$6.2 million for a season, too much reliance on high-definition, high-speed cameras, and in some high-end events, the effect is not as expected, etc. [3]. The current mainstream VAR technology, its accuracy and efficiency are related to the number of cameras, the frame rate of the cameras, the placement position and other factors. In addition, a human assistant needs to be configured, so the effect is also related to factors such as the proficiency of the human assistant's operation or subjective judgment. According to FIFA's intelligence, each club in the UK will pay $1.38 million for VAR in 2022, and many clubs say this will be a huge burden. This is still true for the developed UK, and more serious for some developing countries. The potential drawback for league football in developing countries is that the high cost of using the system may restrict

DOI: 10.24507/icicelb.14.08.803

access to the technology, thereby, widening the gap between the leagues that can afford it and those that cannot. Licensing fees contribute to these costs. The rejection of VAR by twelve of the twenty clubs participating in the Campeonato Brasileiro Serie A (Brazilian top division) in February 2018, highlights this problem. Through Sun and Yi's research [4], we know that the probability of fouls occurring in different areas is different, and the probability of misjudgment is also different. Therefore, if we focus on monitoring those areas that are prone to fouls and misjudgments, we can greatly improve the use effect of video referees under the premise of controlling costs. Based on Jiang and Lin's research [5] on the use cases of VAR systems at the 2018 World Cup in Russia, we found several problems with existing VAR systems. The first point is the lag. It takes time for the video picture to pass the VAR judgment. After the judgment is completed, the auxiliary referee needs to check and check. After the check is correct, it needs to be reported to the chief referee. Finally, the chief referee decides whether to change the penalty. However, the situation on the field is changing rapidly, and this lag may bring more controversy to the game, and also seriously affect the audience's perception. The second point is subjectivity, because the final decision is still made by the referee, so if the referee does not trust VAR, the effect of VAR will be greatly weakened. In the 2018 World Cup in Russia, there were 3 cases of VAR making a penalty, the auxiliary referee verified that it was correct, but the main referee still maintained the original judgment. Therefore, because the existing VAR has the above problems, we want to design a referee system with low cost, and only focus on key areas, less dependence on people, and simple operation. We believe that such a system will be better promoted to underdeveloped countries, to some small events and campus events, helping to improve the fairness of the event and reduce the difficulty of refereeing. In this paper, we propose methods to solve some practical problems through human body recognition technology, such as algorithms to assist referees in making judgments in simple environments such as campuses, and clarify the classification rules and advantages and disadvantages of different algorithms, hypothesized conjectures and proposed implementation methods. Now, many research teams are using motion capture systems to analyze and judge sports games or other behaviors. Maria's research team captures and analyzes parameters such as trajectory and speed of tennis players and their rackets to gain a more data-driven understanding of their performance [6]. Skublewska-Paszkowska's team uses high-precision and high-cost systems such as Xsence motion capture system and Vicon system to conduct experiments, and can obtain accurate data. However, due to its characteristics such as the need to install a large number of supporting equipment in a specific place, it is not suitable for the scene of this paper. Similarly, Zhang's research team used GoogleNet network to do action recognition and analysis in table tennis [7]. Zhang's team improved on the basis of GoogleNet Network and proposed Dual-stream New GoogleNet Network. The network has significantly improved the detection accuracy and generalization ability of table tennis games, but the detection targets of table tennis games are small and the degree of crowding is low. Therefore, the recognition performance of the network in multi-person events such as football games is still unknown. Ma's research team uses the Alphapose algorithm to identify the fall of the elderly and issue early warnings to save the life of the elderly [8]. Ma's team used Alphapose for recognition and achieved good results, so we decided to apply the algorithm to small football events based on Ma's research, and we can expect good results. This paper creatively proposes to use algorithms such as Alphapose to make a low-cost, low-latency, high-automatic video-assisted referee suitable for campus and small events, which can improve the fairness of the game and the viewing experience of the audience, and reduce the referee's burden. This paper is organized as follows. In Section 2, the problem considered in this paper is described. In Section 3, we classify the human key point detection algorithm and explain their respective advantages, focusing on the algorithms used by Alphapose. In Section 4, we introduce target detection and human recognition algorithm based on Convolutional Neural Network (CNN), and focus on the advantages of YOLOv3 in target detection. In Section 5, we introduce the specific implementation method. Section 6 gives concluding remarks.

2. **Problem Formulation.** How to improve the fairness of the referee's penalty, in addition to improving the professional quality of the referee itself, can also improve the accuracy of the penalty through some auxiliary means. We thought of using cameras and human body recognition technology, and how to realize multi-person action recognition and judgment in simple scenes is the problem we need to solve. In the multi-person condition, it is difficult to quickly and accurately identify the human frame and calculate the position of the bones. At the same time, we also need the algorithm to be able to recognize the ball, so that we can judge by the relative position of the person and the ball. In a scenario involving multiple individuals, it is crucial to determine whether the observed object is obstructed by others. Additionally, we need to address the question of how our system can continue to fulfill the role of an auxiliary video referee when our target is blocked. In small events such as campuses, the cameras are usually not very high-end, which makes the definition and frame rate of the picture not very high. How to identify accurately and quickly in this case is also one of the issues to be considered.

To address these issues, we propose a specific algorithm for multi-object recognition in simple scenes using human recognition technology.

3. **Human Key Point Detection Algorithm.** Human recognition technology can be divided into single-target recognition and multi-target recognition by identifying the number of targets. Due to the superiority of Alphapose in the field of multi-target recognition, this article uses Alphapose for motion detection [9]. Alphapose uses Single-Person Pose Estimator (SPPE) to extract specific bone information from the recognition frame. Different recognition frames can be generated by changing the size of the Positioning Box Accuracy (PBA) value. Each frame operates independently to generate different bone recognition lines, and the optimal recognition line is selected as the final result output [10].

Alphapose detects key points such as nose, eyes, shoulders, knees, and elbows from the final output, and uses deep learning and image processing techniques to detect the location and category of human key points in the image, and mark them. By connecting the points correctly, the human pose can be estimated [11, 12]. Human key point detection algorithms are mainly divided into two categories based on different feature extraction methods: traditional skeleton key point detection algorithm and deep learning-based skeleton key point detection algorithm [13, 14]. According to the difference in dimensions, it can be divided into 2D human pose estimation and 3D human pose estimation [15]. Compared with 3D human pose estimation, 2D human pose estimation has the advantages of simple sampling and simple data processing. This paper chooses to use the 2D human pose estimation method for research. 2D human pose estimation mainly includes two aspects: Single Person Skeleton Key Points Detection (SPSKPD) and Multi-Person Skeleton Key Points Detection (MPSKPD). Among them, MPSKPD evolved from SPSKPD, and MPSKPD mainly has two research directions: top-down human skeleton keypoint detection algorithm and bottom-up human skeleton keypoint detection algorithm. The Alphapose algorithm adopts the Regional Multi-person Pose Estimation (RMPE) framework. The basic structure is shown in Figure 1, which is a top-down detection method. It mainly consists of the following three components [10].

1) Symmetric Spatial Transformer Network (SSTN)
2) Parametric Pose Non Maximum Suppression (PP-NMS)
3) Pose Guided Proposals Generator (PGPG)

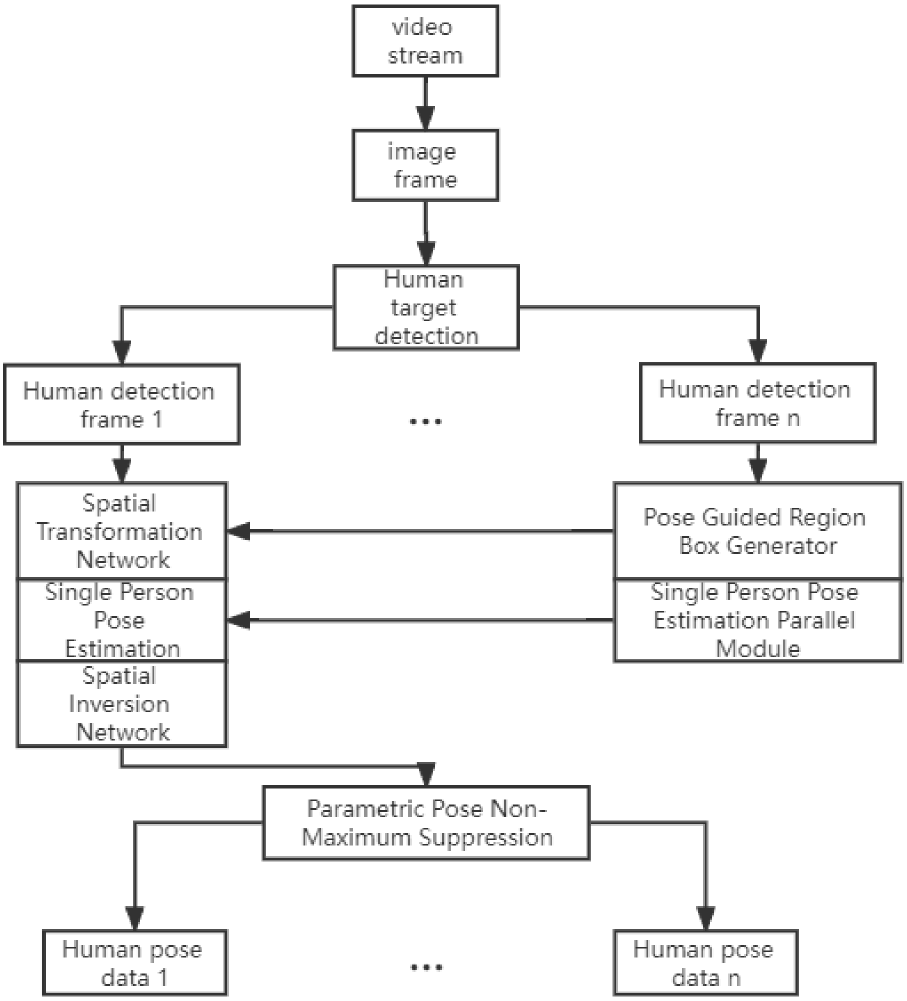

video stream
image frame
Human target detection
Human detection frame 1
...
Human detection frame n
Spatial Transformation Network
Single Person Pose Estimation
Spatial Inversion Network
Pose Guided Region Box Generator
Single Person Pose Estimation Parallel Module
Parametric Pose Non-Maximum Suppression
Human pose data 1
...
Human pose data n

FIGURE 1. Processing flow of RMPE components

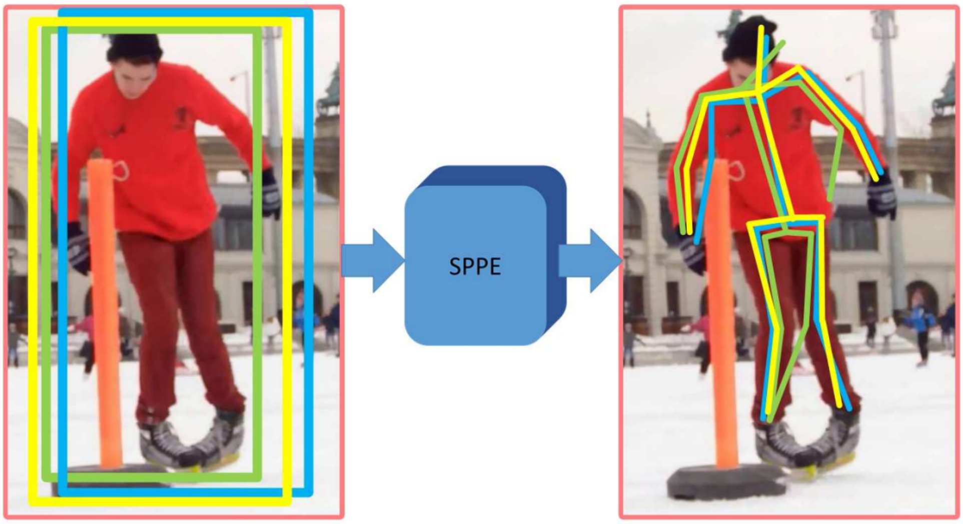


FIGURE 2. The SPPE mode

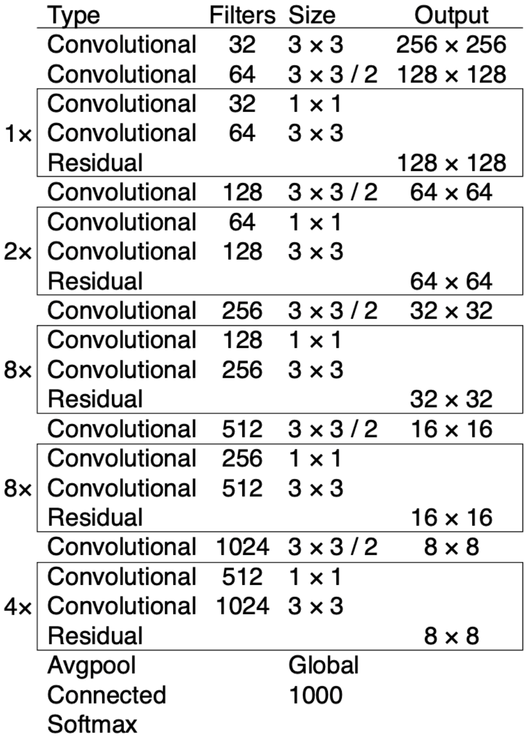
As shown in Figure 1, SSTN includes Spatial Transformer Networks (STN), SPPE and Spatial De-Transformer Network (SDTN). Among them, SDTN is the inverse transformation of STN. Therefore, SSTN is symmetrical. STN accepts the segmented human detection frame. SDTN generates pose proposals. In Figure 2, except for the frame 1, the parallel SPPE modules used by other frame locks are only STN and SPPE. All layer parameters of parallel SPPE are fixed in the training phase. The parallel SPPE branch is compared with the annotation of the real pose, and the pose error of the center position is back-propagated to the STN module. If the pose of the STN is not centered, the parallel SPPE backpropagation will have a large error. Through back-propagation, it helps the STN to focus on the correct area, so as to achieve high-quality human body area extraction. The RMPE framework has good performance in the field of multi-person recognition, and can hit 76.7 [mAP] on the multi-person data set MP2 [16].

4. **Target Detection Algorithm.** In this section, we explain target detection algorithm.

Although we use the Alphapose algorithm to determine the human skeleton feature points, we still need an algorithm to extract the human frame. Most modern human object detection algorithms are based on Convolutional Neural Network (CNN). Unlike conventional neural networks, the neurons in each layer of the CNN network are arranged in three dimensions: width, height and depth [17]. Width and height are used to describe two-dimensional data. The depth of the entire network refers to the number of layers of the network. If the dimension of the input data is $32 \times 32 \times 3$, then after passing through the CNN, the dimension of the final output layer is $1 \times 1 \times 10$. This is because the CNN compresses the image into a vector containing various ratings in the last part. CNN is mainly composed of these types of layers: input layer, convolutional layer, ReLU layer, pooling layer and fully connected layer. By stacking these layers, a complete convolutional neural network can be constructed. After passing through a convolutional layer, the image becomes abstracted to a feature map. The neurons in the layer will only be connected to a small area in the previous layer instead of being fully connected, and the advantage of this is that the number of neurons can be reduced. For instance, a fully connected layer for a (small) image of size 100 × 100 has 10,000 weights for each neuron in the second layer. Instead, convolution reduces the number of free parameters, allowing the network to be deeper. For example, regardless of image size, using a 5 × 5 tiling region, each with the same shared weights, requires only 25 learnable parameters. A CNN is a feed-forward artificial neural network based on the visual cortex. A CNN plays a powerful role in recognizing objects by extracting the features in images [18]. YOLO is a CNN for performing object detection in real time. CNNs are classifier-based systems that can process input images as structured arrays of data and identify patterns between them (view image below). YOLO has the advantage of being much faster than other networks and still maintains accuracy. Due to the excellent performance on the COCO dataset, we plan to use the YOLOv3 algorithm for target detection.

The YOLOv3 algorithm is the third-generation algorithm of YOLO. The YOLOv1 algorithm is characterized by real time, which has the advantage of less false detection and fast speed [19]. However, the disadvantage is that when the center points of different objects of multiple categories fall within the same grid, the classification loss is ambiguous and the positioning accuracy has a big gap compared to faster-RCNN. YOLOv2 is improved on the basis of v1, which is faster than the SSD of the same period. v3 improves the structure and training skills of v2, and improves the accuracy without increasing the inference time.

The author of YOLO regards the target detection problem as a regression problem. First, the input image is divided into S × S grids. If the center point of the target frame is in the grid, then this grid is responsible for predicting the target. This network is mainly composed of a series of 1×1 and 3×3 convolutional layers, with a total of 53 layers. Also known as Darknet-53, the specific structure is shown in Figure 3. Each grid will predict bounding box, confidence and class probability map. The bounding box contains four values: x, y, w, h where (x, y) represents the center point of the prediction box, w, h represent the width and height of the prediction box. Confidence indicates the possibility



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

FIGURE 3. The structure of Darknet-53

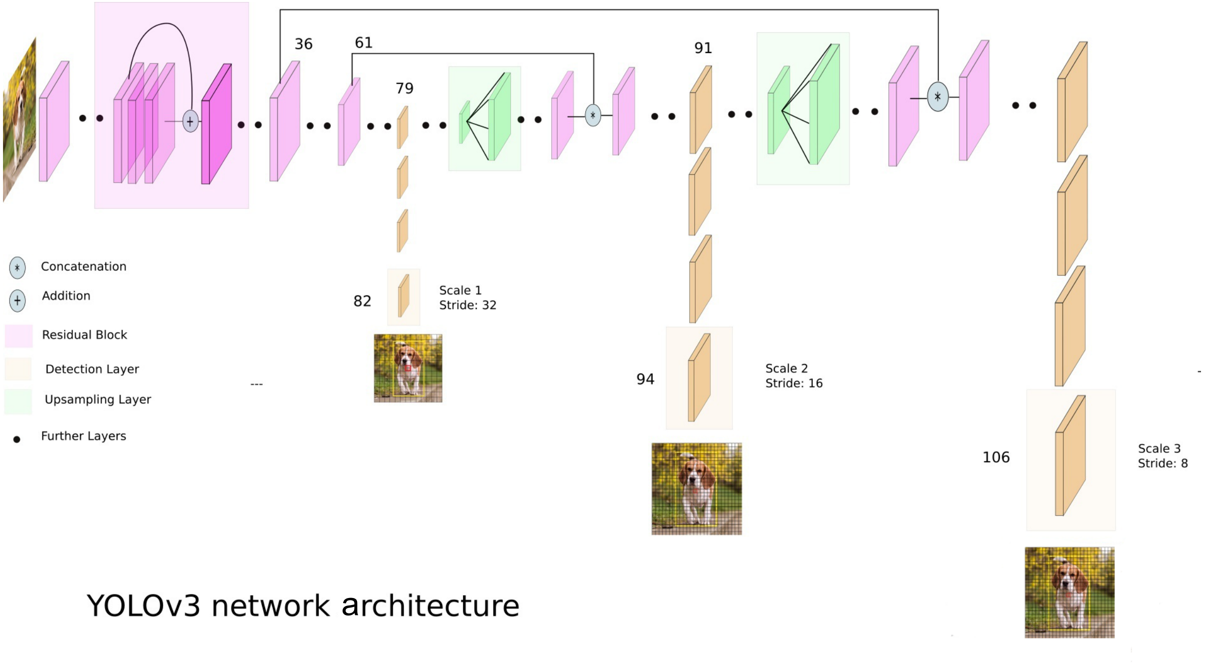


FIGURE 4. (color online) The residual module

that the prediction box contains the target, and the true value during training is the IoU of the prediction box and the true value box. The class probability map indicates the confidence of the class to which this object belongs. As can be seen from Figure 3, Darknet-53 consists of convolutional layers and residual layers. The overall structure of YOLOv3 is shown in Figure 4. From the figure we can see that, YOLOv3 predicts objects of size 8 times, 16 times and 32 times on the feature maps of 3 scales, respectively. That is to say, if our input is 416 × 416, the size of the feature map used in YOLOv3 prediction is 52×52, 32×32 and 13×13. For the first scale, YOLOv3 downsamples the input to 13×13 and predicts at layer 82. At this time, the size of the predicted output 3-dimensional Tensor is $13 \times 13 \times 255$. After that, YOLOv3 takes the feature map from layer 79, then applies a convolutional layer for channel compression, and then upsamples it by a factor of 2 to a size of 26×26. Then, the feature map is concat with the feature map of layer 61. Finally, the feature map after concat is further extracted through several convolutional layers, until the 94th layer is used as the feature map for the second scale detection. The size of the 3D Tensor of the second scale prediction output is $26 \times 26 \times 255$. For the third scale, repeat the above operation. That is, the feature map of the 91st layer is first followed by a convolutional layer for channel compression, then upsampled by 2 times, the size is 52×52, and then the feature map of the 36th layer is concat operation. Then it is several layers of convolution operations, the final prediction layer is completed in 106 layers, and the size of the generated 3D Tensor is $52 \times 52 \times 255$. In summary, YOLOv3 performs detection on feature maps of 3 different scales, so if we input an image of size 416×416, it will produce 3 different output shape Tensors, $13 \times 13 \times 255$, $26 \times 26 \times 255$ and $52 \times 52 \times 255$. The structure of the residual module is shown in Figure 5. The most notable feature of the residual module is the use of a shortcut mechanism to alleviate the problem of vanishing gradients caused by increasing the depth in the neural network, thereby making the neural network easier to train. It mainly uses identity mapping to establish connection between input and output.

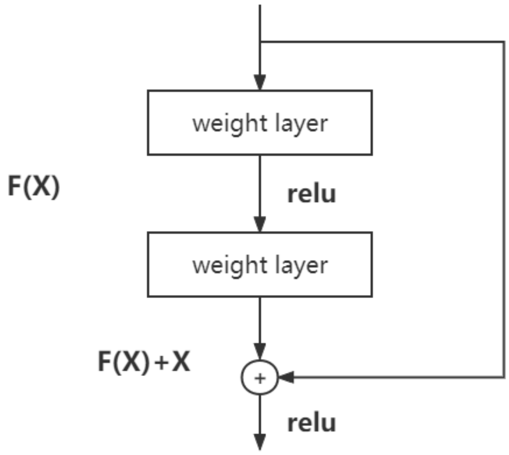


Figure 5. The residual module

YOLO will predict and output on three feature maps of 8 times, 16 times and 32 times, respectively. For CNN, as the CNN extracts the information of the input image, the number of feature maps, that is, the number of convolution kernels, will increase. The consequence of this is that the spatial information will be reduced and the extracted features will become more abstract. However, YOLO greatly optimizes this, reducing the dimension of the Region of Interest (ROI) from $[32, 32, 3]$ to 85 dimensions. The first four dimensions of the feature vector represent the candidate frame information, the middle dimension represents the probability of judging whether the object exists, and the next 80 dimensions represent the classification probability information of 80 categories [19]. YOLOv3 can quickly and accurately detect the bounding box of moving objects, and the effect is shown in Figure 6.

5. **Experimental Process Conception.** First, configuring the environment, the experimental environment is shown in Table 1. After configuring the environment, download and install tools such as Alphapose or RMPE. Then download YOLOv3 detector and put it under the detector/yolo/data folder. Download the pose keypoint detection model and

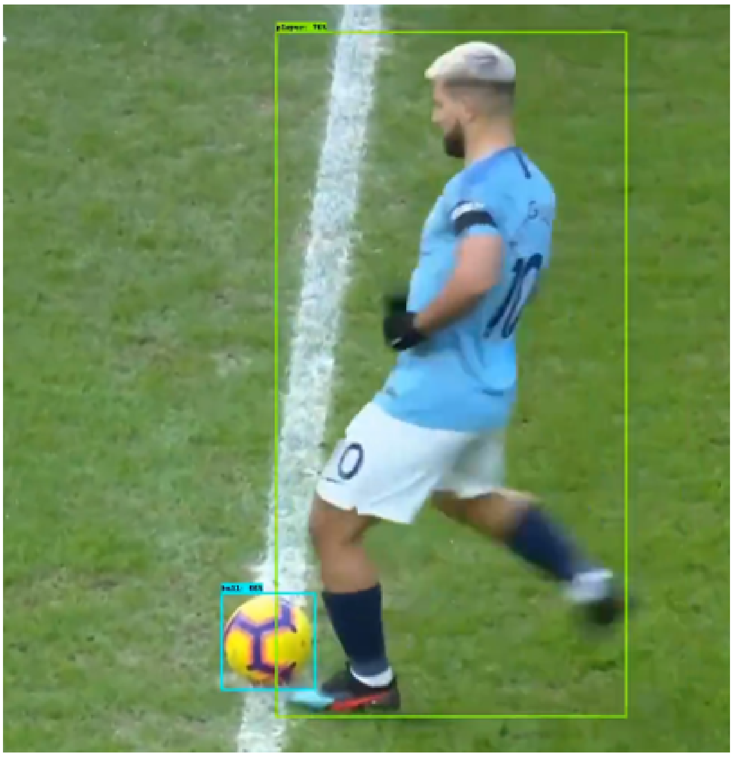


Figure 6. The result of bounding box recognition with YOLOv3

Table 1. Lab environment

OS	Linux
Linux version	Ubuntu 20.04
Cuda version	11.6
GPU	GTX3060
CPU	R7 5700G

the pedestrian re-identification model. Finally, put pictures or videos in the demo folder for identification. Alphapose can do

1) Identify the picture of a single football match;
2) Identify a video of a football match;
3) Use high-definition cameras to identify an ongoing football match.

Alphapose has three powerful functions built in, which are used to identify pictures, videos and live broadcasts. We call Function 1 image recognition, Function 2 video recognition and Function 3 live video stream identification, respectively. Function 1 and Function 2 can be used to check the correctness of the penalty after the game, Function 3 can provide guidance for referee decisions in ongoing matches.

Figure 7 shows the effect of Alphapose's recognition of skiers' movements. It can be seen that the recognition effect is still very good even in the case of multiple people. Through testing, it is found that Alphapose has excellent performance. On the COCO test-dev 2015 dataset, as shown in Table 2, the AP (Average Precision) is better than both OpenPose (CMU-Pose) and Detectron (Mask R-CNN) algorithms. The Alphapose is 10 mAP higher than the previous algorithm RMPE and can reach 20fps, which is a cross-generational improvement. As shown in Table 3, on the MSCOCO dataset, Alphapose also outperforms OpenPose and Mask R-CNN in both fps and mAP. In Figure 7, we can clearly see the bones of those identified players. When the distance between the bones of the two players in a specific part is below a threshold, by recognizing the edge profile of

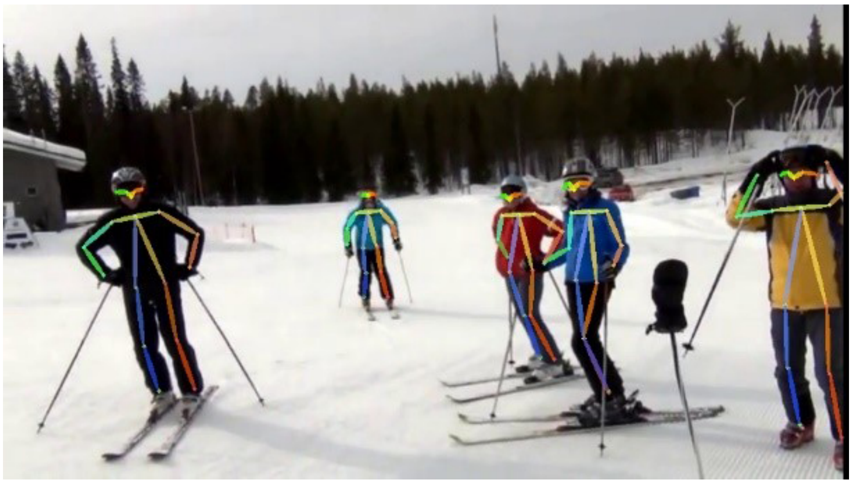


Figure 7. Alphapose body recognition example

Table 2. Performance of several human recognition algorithms on COCO test-dev 2015

Method	AP @0.5:0.95	AP @0.5	AP @0.75	AP medium	AP large
OpenPose	61.8	84.9	67.5	57.1	68.2
Mask R-CNN	67.0	88.0	73.1	62.2	75.6
Alphapose	73.3	89.2	79.1	69.0	78.6

Table 3. Performance of several human recognition algorithms on MSCOCO dataset

Algorithm	FPS	mAP
OpenPose	10	61.8
Mask R-CNN	5	67.0
Alphapose	20	73.3

the player, a good judgement of the presence of a contact foul can be made. However, the judgment from one angle is inaccurate, so we may need to synthesize information from multiple angles to make a comprehensive judgment. For example, if all cameras in different directions have detected a collision between the key points of the ball and the hand, and the trajectory of the ball has changed significantly, it may indicate that a handball violation has occurred. In order to reduce the possibility of misjudgment, we need to make a scoring system. When a key point contact occurs or the ball trajectory changes, the score will be increased, and the score will be determined according to factors such as the number of contact cameras and the magnitude of the trajectory change. How much to judge a violation if the score exceeds the threshold.

6. **Conclusion.**

In this paper, we have proposed methods to solve some practical problems through human body recognition technology, such as algorithms to assist referees in making judgments in simple environments such as campuses, and have clarified the classification rules and advantages and disadvantages of different algorithms, hypothesized conjectures and have proposed implementation methods. In the following research, we will focus on building the scoring system for foul judgment mentioned above, so as to realize the foul judgment function in simple scenarios. At the same time, we will optimize the algorithm to improve the speed and accuracy of judgment in simple scenarios.

REFERENCES

[1] H. Koyama, N. Kida and K. Ookado, Analysis of eye line of referee in kendo game, *Proc. of Mechanical Engineering Congress Japan*, J2010604, DOI: 10.1299/jsmemecj.2018.J2010604, 2018.

[2] K. Abe, Outcomes of VAR and investigation of incidents: The case of the 2018 FIFA World Cup, *Journal of Gakushuin University Research Institute for Humanities*, pp.247-256, 2019.

[3] J. Xu and K. Tasaka, Keep your eye on the ball: Detection of kicking motions in multi-view 4K soccer videos, *ITE Transactions on Media Technology and Applications*, pp.81-88, 1985.

[4] L. Sun and H. Yi, A comparative analysis of the use of VAR in football video assistant referees at home and abroad, *Contemporary Sports Technology*, pp.217-219, 2020.

[5] X. Jiang and C. Lin, Analysis of VAR usage in the 2018 FIFA World Cup Russia, *Journal of Sanming University*, no.2, pp.95-100, 2019.

[6] M. Skublewska-Paszkowska, E. Lukasik and J. Smolka, Algorithms for tennis racket analysis based on motion data, *Advances in Science and Technology Research Journal*, pp.255-262, 2016.

[7] A. Zhang and H. Yu, Space-time dual-stream table tennis action recognition based on improved GoogLeNet network, *Computer Knowledge and Technology*, pp.78-80, 2021.

[8] J. Ma, H. Lei and M. Chen, Algorithm for detecting fall behavior of the elderly based on AlphaPose optimization model, *Computer Applications*, vol.42, no.1, pp.294-301, 2022.

[9] T. Haruyama, S. Takahashi, T. Ogawa and M. Haseyama, Multimodal important scene detection in far-view soccer videos based on single deep neural architecture, *ITE Transactions on Media Technology and Applications*, pp.89-99, 2020.

[10] H.-S. Fang, S. Xie, Y.-W. Tai and C. Lu, RMPE: Regional multi-person pose estimation, *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] J. Shin and S. Ozawa, A study on motion analysis of an artistic gymnastics by using dynamic image processing, *Journal of the Japan Society for Precision Engineering*, pp.669-673, 2009.

[12] D. Li, D. Zhao and M. Zhi, Research on human behavior recognition based on deep learning, *Journal of Frontiers of Computer Science Technology*, pp.1-15, 2022.

[13] G. Bao, D. Li and Y. Mei, Key frames extraction based on optical-flow and mutual information entropy, *Journal of Physics: Conference Series*, 012112, 2020.

[14] H. Miyama and S. Inoue, Moving object tracking by function distributed architecture and feedback operation, *ITE Technical Report*, pp.61-68, 1997.

[15] L. Ma, S. Lian, S. Wang, W. Meng, J. Xiao and X. Zhang, A practical framework of multi-person 3D human pose estimation with a single RGB camera, *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp.420-421, 2020.

[16] M. Sun, P. Kohli and J. Shotton, Conditional regression forests for human pose estimation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3394-3401, 2012.

[17] S. Hoshino and K. Niimura, Optical flow for real-time human detection and action recognition based on CNN classifiers, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol.39, pp.735-742, 2019.

[18] B. Hao, S. Park and D. Kang, Research on pedestrian detection based on Faster R-CNN and hippocampal neural network, *2018 10th International Conference on Ubiquitous and Future Networks (ICUFN)*, pp.748-752, 2018.

[19] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, *arXiv.org*, arXiv: 1804.02767, 2018.