Computational Analysis of Table Tennis Matches from Real-Time Videos Using Deep Learning

Hong Zhou, Minh Nguyen, Wei Qi Yan
Auckland University of Technology, Auckland, New Zealand

Abstract. In this paper, utilizing a multiscale training dataset, YOLOv8 demonstrates rapid inference capabilities and exceptional accuracy in detecting visual objects, particularly smaller ones. This outperforms transformer-based deep learning models, makes it a leading algorithm in its domain. Typically, the efficacy of visual object detection is gauged by using pre-trained models based on augmented datasets. Yet, for specific situations like table tennis matches and coaching sessions, fine-tuning is essential. Challenges in these scenarios include the rapid ball movement, color, light conditions, and bright reflections caused by intense illumination. In this paper, we introduce a motion-centric algorithm to the YOLOv8 model, aiming to boost the accuracy in predicting ball trajectories, landing spots, and ball velocity within the context of table tennis. Our adapted model not only enhances the real-time applications in sports coaching but also showcases potential for applications in other fast-paced environments. The experimental results indicate an improvement in detection rates and reduced false positives.

Keywords: YOLOv8 · Moving balls · DETR · Image pre-process · Image post-process · Background subtraction · Deep learning

1 Introduction

Deep learning has gained applications in sports competitions, particularly in tasks such as determining the placement of balls in table tennis. This shows inherent challenges in table tennis, such as the diminutive size and subtle texture patterns of table tennis balls. Compared to other sports, table tennis balls can be hard to distinguish from background textures, complicating the process of determining the landing points and velocities.

In addressing the complexities associated with detecting and identifying visual objects in fast-moving environments such as table tennis, the selection of the most optimal model emerges as an indispensable step. Currently, deep learning predominantly features in two mainstream ways: YOLOv8 algorithm [1] and transformer-based algorithms [2] for computer vision tasks. While the solutions for the dynamic and real-time requirements of table tennis training and actual competitions, the speed of real-time object detection becomes paramount. This elevates inference time to a critical determinant in algorithm selection process.

Among the contenders, YOLOv8 distinctively stands out. Not only is it more streamlined, but it also boasts a markedly rapid inference time, making it a preferable choice

over many transformer-based algorithms. To provide a holistic understanding of YOLOv8 efficacy, this paper embarks on a comparative analysis with DETR (End-to-End Object Detection with Transformers), a flagship representation of transformer-based algorithms. This comparison underscores the relative advantages of YOLOv8 model, particularly spotlighting its superior performance in inference speed and proficiency in detecting small objects. Beyond just its speed, the YOLOv8 model is underpinned by a cutting-edge architectural design coupled with avant-garde training methodologies. These components synergize to achieve heightened precision in localizing and recognizing diminutive objects within images, even in the most challenging scenarios. This positions YOLOv8 as not just an alternative, but potentially the future standard in object detection for real-time applications.

The solid color of table tennis balls can cause them to be mistaken for light sources. We observed many instances where background items were misidentified as balls, negatively impacting detection rates. To address this, we've integrated a module that focuses on motion patterns in constancy of the balls. This module employs background subtraction to differentiate between static background and moving foreground elements, based on the parameters detailed in this paper. This approach enhances density estimation, clustering similar data together. Once the background is removed from the video sequence, the ball path is evident through aiding the YOLOv8 model in extracting visual cues and predicting outcomes.

High-speed cameras can potentially mitigate motion blur challenges faced when capturing swift moving balls in table tennis. Still, the inference time of YOLOv8 model struggles to match the speed of digital cameras. To ensure proper detection, the training dataset must accommodate various ball shapes, including the distortions from motion blur. Adjusting the camera speed in frame per second can provide a more diverse training dataset for the model.

Measuring the actual velocity of a table tennis ball in the entails determining its three-dimensional path, emphasizing on the importance of object distance. While Lidar can detect small and reflective objects like the table tennis balls, inaccuracies can arise due to the laser interaction with such surfaces. Camera calibration presents a more reliable method for determining the ball depth across successive frames.

Properly pinpointing where the ball lands on the table mandates precise detection, especially on the table surface. Traditional evaluation methods, such as comparing predicted bounding boxes with ground truth boxes, may not be sufficient. Thus, we propose a novel evaluation method focusing on the landing point of the balls on the table.

This paper systematically delves into literature reviews, methodologies, outcomes, discussions, and conclusions. It comprehensively covers model structures, experimental strategies, and algorithm deployment.

2 Literature Review

The velocity that a flying ball in table tennis games is the specific feature. This fast-moving object which generated by the players' explosive power of swinging the bat and hitting a ball is a challenge in computer vision. By considering the speed factors,

the landing spot where a player hits the ball on the table is also an evaluation of playing skills.

A prior study highlighted the formidable challenge of ball detection within the realm of computer vision, attributed to diminutive size and swift motion of the balls [3], even YOLOv8 model struggles with a variation of aspect ratio and accurately detects balls due to fast motion. While an anchor-free approach was proposed to counter this issue during the evolution of YOLO models. This challenge in detecting such balls is still difficult as the state-of-the-art algorithm. Nonetheless, the limitations may not only stem from false positives by using anchor-free algorithms, but the small size of sports balls could also cause to a significant influence.

A valuable reference was dedicated to track moving or airborne objects. In 2022, a method employing LSTM in deep learning and simple physical motion models corrected deviations, through establishing a binocular vision-based trajectory extraction system for table tennis that relies on digital cameras [4]. The visual feature extraction was completed by using MobileNet and SSD models, a compromise between resource-constrained environments and accuracy. Nevertheless, it falls compared to the pyramid feature network in YOLOv8 architecture, particularly for challenging datasets and small visual objects.

After reviewed the video footages of 2017 Summer Universiade Men's Singles Final, persisting in achieving precise recognition and positioning of high-speed, small balls was considered a challenge [5]. The TrackNet model, built upon deep learning, can identify balls from single frames with blurred images and lingering trails, even unable to be seen from a visual perspective. However, the performance of TrackNet model heavily hinges on training data it encounters, potentially faltering if exposed to visual objects or environments deviating significantly from the training data.

VAR (i.e., Video Assistant Referee) was available in the 2018 FIFA World Cup, volleyball matches, and fencing competitions. Conversely, it has not been applicable in table tennis competitions due to the exceptional speeds of balls up to 112.5 kilometers per hour [6]. As a widely participated sport with 800 million table tennis players globally, it laid the foundation for popularity ranking at the Olympics. Tracking and detecting table tennis are anything but routine. Employing VAR introduces the risk of misjudgment constrained by the ball's incredible speed.

Apart from overcoming the challenges associated with tracking and detecting table tennis, we have to face the problems related to the relationship between training datasets and accuracy enhancement. A group of models struggle to achieve the officially announced accuracy, with actual detection results falling short of expectations. Fast or erratic object motion causes motion blur [7], making it difficult to comprehensively cover training datasets and assess detection outcomes.

An end-to-end BFAN (i.e., Blur-aid Feature Aggregation Network) for visual object detection has been proposed [8]. However, the application of this approach seems unsuitable for table tennis due to its requirement for multiscale feature training datasets. Deblurring may restore clarity to the balls in consecutive frames, yet distinguishing blurred foreground from background poses a significant challenge.

Optimizing the predicted bounding box scale might offer a solution. This entails learning scale features from a few samples, as demonstrated by using MSNN (i.e., MultiScale Meta-relational Network) [9]. MSNN enhanced the generalization capability of the proposed model for measurements and improving classification accuracy without necessitating model-independent meta-learning algorithms. While the dataset yielded positive results, further research work was required to fine-tune meta-learning methods for improving the performance on other datasets.

In order to calculate the ball speed of a table tennis using computer vision, it is necessary to find the depth of scenes of table tennis in digital images. The movement of balls of table tennis may be perpendicular to camera lens, which requires at least two fixed cameras to synchronously record from different angles so as to avoid the ball in table tennis being considered as not moving during two consecutive frames. A few of camera APIs provide timestamp information for captured video frames satisfying stereo vision. The frames from different cameras can be aligned using these timestamps. This approach might require prudently handling of the timestamps and proper synchronization logic.

A stereo camera installed on a robot has been studied for tracking table tennis balls after being synchronized [9]. It explores a method that captures and processes stereo images of the ball motion and analyses the disparities between corresponding points in the stereo images, they determine the ball 3D position in space. This method focuses on image synthesis after asynchronous cameras captured images, even if only one camera's frame rate is known. However, this method increases the processing interval time for each frame, which seems to significantly increase the detection time of motion-based YOLOv8 algorithm.

In this experiment, replacing the image information captured by using camera with the image information obtained by using auxiliary cameras only occurs when the table tennis ball captured in two consecutive frames in the main camera are in the same position. In other words, this calculation method only changes the data source from the main camera to an auxiliary camera, without increasing the computational workload under limited computing resources.

In summary, diversifying training dataset scales, deblurring table tennis affected by motion blur, and employing the multiscale meta-relational network appear as viable avenues for investigation. The focus of this paper is on dataset scale diversity while deblurring methods will be explored in subsequent research endeavors.

3 Methodology

3.1 Customed training dataset

The scene of table tennis is specific with a moving small ball that is different from the released training datasets. Thus, a customed training dataset needs to be tailored resulting in better performance. However, it requires effort in terms of data collection, annotation, and quality control. Fortunately, a huge number of parameters can be utilized from pre-trained models through transfer learning. A great number of factors will affect

establishing a customed training dataset. The approach how to collect enough data samples is the first challenge.

According to real scene of table tennis competitions and training, the table occupying the entire width of the video frame seems to be the dieal position with the appropriate angle, which can maximize the size of table as the target detected in the frame without missing any landing spots on the table. This is also conducive for prediction using YOLOv8 models. The real-time video footages captured under these conditions can serve as the main source of images in the training dataset. In addition, the scale of data needs to be enriched through methods based on computer vision to improve data diversity, such as random resizing. The shape of balls in table tennis is simulated at different depths using the frames by setting random scale factors.

On the other hand, the fast-moving object leads to motion blur after captured by a camera, even if the camera has 120 Hz plus the fastest inference time. It is easier to obtain the ball shape under this deformation by using a low-speed camera to capture images. Meanwhile, this type of images with motion blur also requires a randomly resizing. Finally, the balls in table tennis games with various textures and colors need to be used for sampling and recording.

3.2 Modelling

Swin transformer [2][10][11] is a hybrid architecture which is good at large-scale image classification tasks that are efficient by using hierarchical windows and local self-attention mechanisms. In contrast, regarding Swing transformers, DETR [2][10] is much versatile and efficient, the primary focus of DETR is on visual object detection. A bipartite matching loss is deployed for a set of prediction tasks for visual object detection. As a result of eliminating the need for anchor-based methods, object classes and locations are directly predicted in a single forward pass. Thus, DETR is selected to determine which one is much suitable for this experiment that will be conducted in a Google Collab virtual environment equipped with a V100 GPU (graphics processing unit) to satisfy the basic requirements of real-time object detection.

After a real-time video was processed and ball position is predicted by using YOLOv8 model, it is obvious to see that two textured regions in the background are recognized as balls in Fig. 1. In this situation, it is almost impossible to calculate the ball speed and landing spots of a ball. The detection results with these errors cannot be filtered by using shapes and colors, even texture. The correct detection rate of a real ball has become the key to visual object detection.

Fig. 1. Original video with the prediction results by using YOLOv8s model

Real-time recordings from table tennis training and gaming contain abundant light spots and reflective patches in the background. MoG (Mixture of Gaussians) is employed for background subtraction to remove light spots and reflective patches in the pre-processing stage before the video as input to be predicted by using YOLOv8 model.

Based on precise location of balls in table tennis games detected by YOLOv8 algorithm from video frames, the velocity of balls can be calculated through the variation of location between two consecutive frames corresponding to the frame rate. A camera with 120 Hz can effectively prevent the disappearance of the table tennis ball in each frame and ensure the surface of table completely exists in the screen, whilst maintaining an angle that allows for landing spot on the table. An auxiliary camera is fixed at a 90-degree angle to the main camera. Once a ball of table tennis moves perpendicular to the main camera lens in two consecutive frames, which will be replaced by the image information captured by the auxiliary camera in two consecutive frames. The velocity of a ball in each two-dimensional space needs to be mapped to real-world 3D coordinates by using camera calibration which depends on the perspective transformation of a black and white chessboard as a reference that was put in the scene on the table. The intrinsic and extrinsic parameters including focal length principal point, position and orientation obtained lead to depth information added into the coordinates of bounding boxes. The instantaneous velocity with spatial direction can be calculated through the mapped spatial displacement and frame rate.

In order to detect the landing spots, we detect the surface of table as shown in Fig.2. Each side of the table is split into nine regions respectively. One table has left side or right side from the camera viewpoint. The landing spots of the table tennis ball on both sides of the table will be continuously recorded and displayed to the players in percentage form based on the number of times the region of table has been hit. The probability of each region hit by the ball can be analyzed for understanding the players' skills in the aftermath of a match.

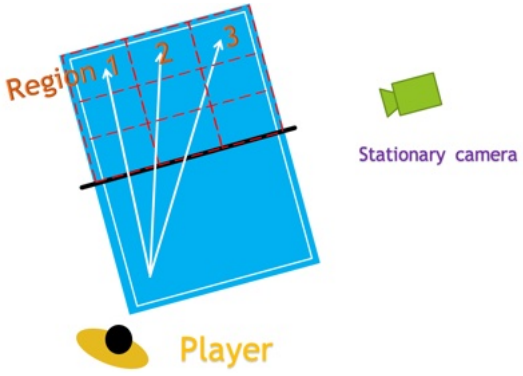

Fig. 2. The sketch of a table division in table tennis

3.3 Methods

The light spots and reflective patches are immovable in a video while using a static camera for ball detection in table tennis games, which provides feasibility for addressing these influencing factors. MoG (Mixture of Gaussians) can subtract background in video sequences based on the pixel intensities in the background. Generally, the initialized approach requires a trade-off between subtle differences in background and computational efficiency. If the pixels represented as Gaussian distribution mixtures are considered as background by using one or more components, it is most possible to be evaluated as background by using MoG.

$$N(\mu, cov) = \frac{1}{\sqrt{2\pi\ det(cov)}} e^{-0.5(\chi-\mu)'(\chi-\mu)inv(cov)} \tag{1}$$

$$P = \sum[weight_i \cdot N(\mu_i,\ cov_i)] \tag{2}$$

The probabilities assigned to each Gaussian component represent a potential class of a Gaussian distribution, such as background. The higher the likelihood, the greater the probability that it belongs to the background. By setting and adjusting the threshold, the accuracy of this classification method can be controlled. Fig. 3 displays the frame of a real-time video that a mask is added to cover the background, the static objects including the light spots and reflective patches are removed after pre-processing. In this experiment, the background subtraction approach does not seem to reduce the accuracy of table tennis ball detection though the color and texture of the ball in each frame is replaced by a white mask.

The results of YOLOv8 prediction involve a 2D tensor of bounding box coordinates. The center point of a table-tennis ball that occurs on the screen is signed as the current box time and coordinates which need to be transferred to real-world coordinates using the perspective transformation with an initialized z-coordinate added as a 2D homogeneous point (x, y, 1). The camera projection matrix combines intrinsic and extrinsic parameters, the inverse matrix is employed to transform points from image coordinates to normalized camera coordinates f_x and f_y are the focal lengths along x-axis and y-axes; c_x and c_y are the optical center coordinates. (X, Y, Z) is the center point coordinates of the ball in the real world.

$$(X, Y, Z) = \begin{bmatrix} 1/f_x & 0 & -c_x/f_x & 0 \\ 0 & 1/f_y & -c_y/f_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot (x,\ y,\ 1) \tag{3}$$

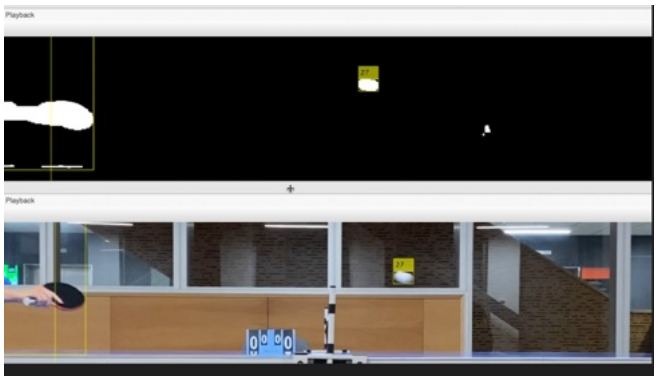

Fig. 3. Separating the moving object from background in an image sequence.

The displacement of a ball between the consecutive frames can be calculated based on coordinate transformation by using eq. (3), the instantaneous velocity will be acquired depending on this time interval, which is the frame rate.

In the grayscale image, corners are at where there are rapid changes in intensity in both the horizontal and vertical directions. Harris corner detection algorithm discovers local intensity variations in the image that are characteristic of corners. Fig. 4 displays the findings of the internal corner on the chessboard placed on the table for camera calibration. The surface of the chessboard is considered perpendicular to *z*-axis in world coordinates and coincident with the plane enclosed by using *x*-axis and *y*-axis. The intrinsic, distortion coefficients, rotation vectors and translation vectors can be computed through the mapping points of the internal corners in the real-world due to the known number of rows and columns in the chessboard and the size of each square in the real world [12].

$$Matrix = \begin{bmatrix} Sum({G_x}^2) & Sum(G_x \cdot G_y) \\ Sum(G_x \cdot G_y) & Sum({G_y}^2) \end{bmatrix} = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \tag{4}$$

$$\lambda = \frac{(A+B) \pm \sqrt{(A+B)^2 - 4(AB - C^2)}}{2} \tag{5}$$

On the table surface, the most significant manifestation of a table tennis ball bouncing after hitting the surface of the table is the change in velocity direction in *y*-axis when the camera and the table are pointing in the same plane. That means, *y*-coordinates for the center point of the table tennis ball change in the vertical axis in consecutive three frames. The function *sign* ($y_m - y_f$) represents the sign of position difference in *y*-axis between the previous two frames consecutively, *sign* ($y_c - y_m$) shows the sign of change of ball positions in *y*-axis between the current frame and the previous one. In eq. (6), if *LS* = -1, the ball hits on the table and then bounces back; or else, the ball of table tennis is considered flying without hitting the table. The bottom of the bounding box for the table tennis ball in the previous frame is compared with the regions to determine where it lands.

Fig. 4. Finding of the corners of a chessboard in an image for camera calibration

$$LS = \begin{cases} not\ hit,\ if\ sign\ (y_m - y_f) \cdot sign(y_c - y_m) = 1 \\ hit,\ if\ \ sign(y_m - y_f) \cdot sign(y_c - y_m) = -1 \end{cases} \tag{6}$$

3.4 Results and Discussions

Fig. 5 is an example after an original image is resized, different resized images will be obtained with 10 random scale factors set as variables for the balls in table tennis games. Fig. 6 demonstrates the resized images with motion blur as an example. This deformation exhibits a variety of forms due to a diversity of directions and velocities of motions, such as rectangles, arches, and shapes approximate to the letter 'v'.

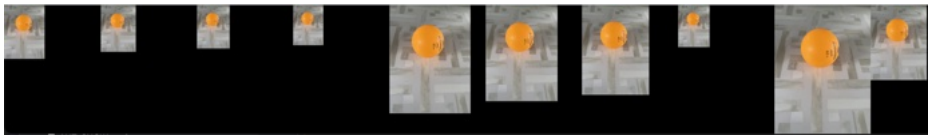

Fig. 5. The example after an original image is resized.

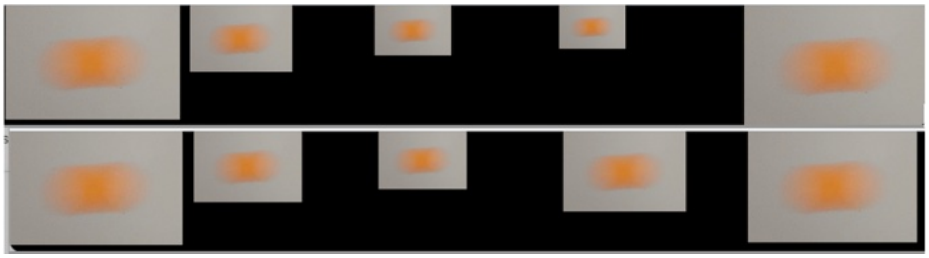

Fig. 6. The example after an original image with motion blur is resized.

The balls for table tennis games with different colors and textures can be detected, Fig. 7 displays five types of colors detected by using YOLOv8 algorithm before the motion-based algorithm is added.

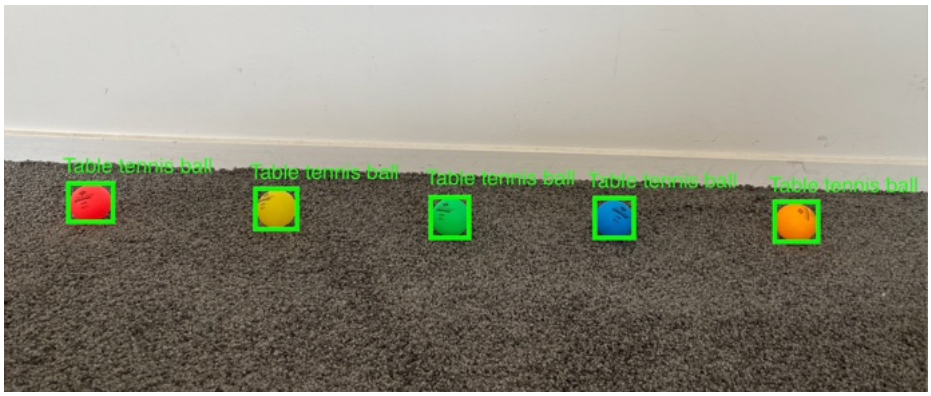

Fig. 7. Five types of colors for the balls in table tennis games are detected by using YOLOv8 model.

The weights and parameters of the pretrained COCO dataset through YOLOv8 model are employed to start with the original 1,774 images captured by 30 Hz and 60 Hz cameras in table tennis training and competition. Accompanied by increasing the amount of random scale factor of resizing to trade-off the influence of the incremental size of the training dataset and the testing dataset on the improvement of AP (Average Precision) based on the prediction of new data, the effectiveness of background subtraction, class matching and circular evaluation, then exploring the optimal solution to terminate the excessive expansion of the training dataset in the case. As a result of illumination, shadows may be generated when the ball approaches the table, and it can be recognized as a ball. In the experiment, although it is not the optimal solution, shadows considered as balls were removed through position-based determination.

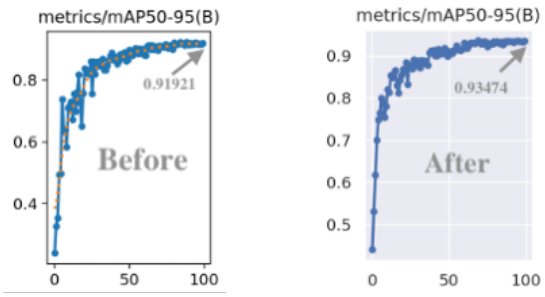

Fig. 8. Comparison of mAP50-95 in 100-th epoch before and after resizing by random scale factor.

The mAP50-95 (i.e., mean Average Precision at 50 and 95 recall) shown in Fig. 8 increases from 0.91921 to 0.93474 after 100th epoch training of motion-based YOLOv8s model that means the model is now better at accurately detecting and localizing objects across a variety of recall levels, 90% of 17,740 images were employed as training datasets, and the remaining images are left for the validation dataset and testing

dataset. The inference time and the detection accuracy of table tennis balls are significant evaluations. Table. 1 illustrates that the inference time of the YOLOv8 algorithm with a shorter inference time compared with the DETR algorithm in the experiment under Google Colab environment equipped with V100 GPU.

Table 1. Comparisons between DETR and YOLOv8

Name	Size(pixel)	Backbone	Inf_time(ms) V100 GPU
DETR	640	ResNet-50	75
YOLOv8m	640	CSPDarknet53	7.2

Fig. 9 displays the table which is automatically divided into nine regions on each side for visual object detection through region segmentation. The bounding boxes of balls touch these regions on the table surface are the landing spots.

(a)  (b)

Fig. 9. In the scene, the table surface is automatically segmented into nine regions on each side. (a) color images with 9 regions on each side (b) Binary images of the table.

Fig. 10 demonstrates the real-time analysis system of table tennis matches. On the left side, it is the video footage of ongoing competition captured by using a 120 Hz stationary camera. The statistics and analysis listed on the right side consist of the instantaneous flying speed of the ball and the percentages of the regions that are hit by the ball on the table. Through this system, both the player and coach can accurately grasp the player's actions that can set training plan for further improvement. Compared with table tennis robot machine which was used as ground truth to serve 50 times at speeds of 20 kilometers per hour, 40 kilometers per hour, and 60 kilometers per hour; The real-time analysis system of table tennis matches was applied for speed measurement, the average accuracy can reach 95%. By manually counting 100 times landing spots of a table tennis ball in a competition, the accuracy of the statistics reaches 99%.

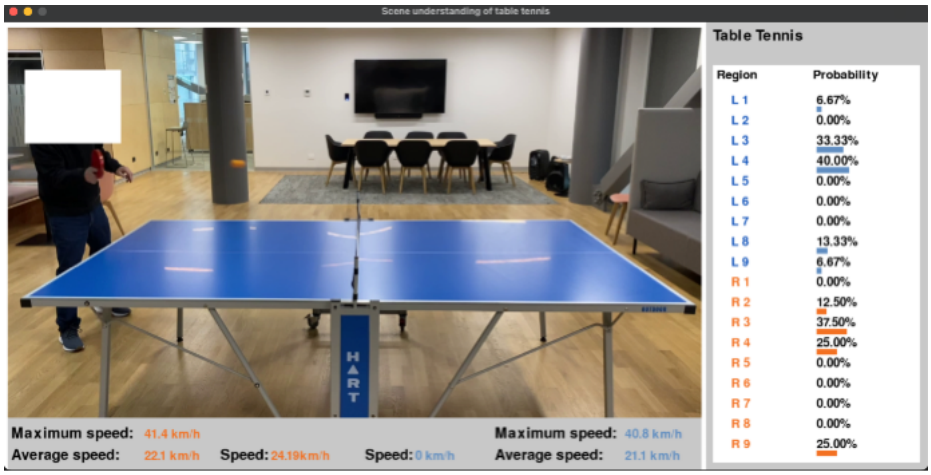

Fig. 10. The interface of real-time analysis of table tennis matches

3.5 Conclusion and Future Work

In this comprehensive study, we delve deep into the fusion of motion-based features and the formidable capabilities of the YOLOv8 model to precisely detect balls in table tennis matches. Our objective extends beyond mere detection by targeting the precise estimation of landing spots and ball velocity. To achieve a detailed, nuanced understanding, we harnessed the capabilities of high-resolution cameras operating at both 30 Hz and 60 Hz. These feeds were then enriched through the adoption of multiscale variation techniques, designed explicitly for data augmentation. This methodological enhancement not only amplified the AP (i. e., Average Precision) but also dramatically curtailed the instances of false positives often instigated by intrusive light reflective interferences.

An innovative inclusion in our research methodology was the deployment of stereoscopic cameras. These cameras are often deployed to capture depth and dimension, which presented a unique advantage. They facilitated the extraction of multiple perspectives and depth information, all while circumventing the typical computational overheads associated with such depth extraction. This strategic utilization paves the way for a precise computation of both the landing spot and ball velocity, leveraging deep learning to decode and interpret real-world video data from table tennis tournaments.

In order to revolutionize table tennis competitions and training regimes by seamlessly integrating motion-centric algorithms, one conspicuous hurdle we encountered was the erroneous recognition of ball shadows as tangible objects, an artifact of the ground subtraction. Although one could potentially discriminate between the actual object and its shadow based on their respective positions during landing spot calculations, we opine that a more holistic solution might lie in the preliminary stages. By refining our pre-processing approaches to systematically eliminate shadows from video frames,

References

1. Xiao, B., Nguyen, M., Yan, W.Q.: Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications, pp.1-18 (2023).
2. Yan, W.: Computational Methods for Deep Learning: Theory, Algorithms, and Implementations, Springer (2023)
3. Tian, B., Zhang, D., Zhang, C.: High-speed tiny tennis ball detection based on deep convolutional neural networks. In International Conference on Anti-counterfeiting, Security, and Identification (ASID), pp. 30–33 (2020).
4. Cai, G, L.: A method for prediction the trajectory of table tennis in multirotation state based on binocular vision. Computational Intelligence and Neuroscience (2022).
5. Huang, Y., Liao, I., Chen, C., İk, T., Peng, W.: TrackNet: A deep learning network for tracking high-speed and tiny objects in sports applications. In IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8 (2019).
6. Moshayedi, A, J., Chen, Z., Liao, L., Li, S.: Kinect based virtual referee for table tennis game: TTV (Table Tennis Var System). In International Conference on Information Science and Control Engineering (ICISCE), pp. 354–359 (2019).
7. Shi, J., Xu, L., Jia, J.: Discriminative blur detection features. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2965–2972 (2014).
8. Wu, Y., Zhang, H., Li, Y., Yang, Y., Yuan, D.: Video object detection guided by object blur evaluation. IEEE Access, vol. 8, pp. 208554–208565 (2020).
9. Zheng, W., Liu, X., Yin, L.: Research on image classification method based on improved multi-scale relational network. J Computer Science, vol. 7, pp. 613 (2012).
10. Yan, W.: Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics, Springer (2019).
11. Xiao, B., Nguyen, M., Yan, W.Q.: Apple ripeness identification from digital images using transformers. Multimedia Tools and Applications, pp.1-15 (2023).
12. Liu, Y., Liu, L.: Accurate real-time ball trajectory estimation with onboard stereo camera system for humanoid ping-pong robot. Robotics and Autonomous Systems, vol. 101, pp. 34–44, (2018).
13. Mehtab, S., Yan, W.: Flexible neural network for fast and accurate road scene perception. Multimedia Tools and Applications (2022)
14. Mehtab, S., Yan, W.: FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. International Conference on Control and Computer Vision (2021)
15. Xiang, Y., Yan, W.: Fast-moving coin recognition using deep learning. Multimedia Tools and Applications (2021)
16. Qi, J., Nguyen, M., Yan, W.: Small visual object detection in smart waste classification using Transformers with deep learning. IVCNZ (2022)
17. Liu, J., Pan, C., Yan, W.: Litter detection from digital images using deep learning. Springer Nature Computer Science (2022)
18. Pan, C., Yan, W.: Object detection based on saturation of visual perception. Multimedia Tools and Applications 79 (27-28), 19925-19944 (2020)
19. Pan, C., Yan, W. A learning-based positive feedback in salient object detection. IVCNZ (2018)
20. Pan, C., Liu, J., Yan, W., Zhou, Y.: Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing (2021)
21. Qi, J., Nguyen, M., Yan, W.: CISO: Co-iteration semi-supervised learning for visual object detection. Multimedia Tools and Applications (2023)