engineering proceedings

MDPI

Proceeding Paper

Computer Vision-Based Tennis Ball Tracking Using You Only Look Once for Training Analytics †

Pei-Jung Lin *, Yu-Tsen Lin, Yong-Liang Lin, Yi-Ping Lee and Shao-Wei Chang

Department of Computer Science and Information Engineering, Feng Chia University, Taichung 40724, Taiwan; linjohnny500@gmail.com (Y.-T.L.); liny20021001@gmail.com (Y.-L.L.); taiwan20170114@gmail.com (Y.-P.L.); bb930610@gmail.com (S.-W.C.)

* Correspondence: linpj@fcu.edu.tw

† Presented at the 7th Eurasia Conference on IoT, Communication and Engineering 2025 (ECICE 2025), Yunlin, Taiwan, 14–16 November 2025.

Abstract

Tennis is an exceptionally fast-paced sport where the ability to return the ball precisely to an opponent's weak zones often determines match outcomes. Although wall practice serves as a fundamental and effective training method, accurately capturing and analyzing the spatial distribution of ball impact points during high-speed rallies remains highly challenging. Leveraging computer vision, we propose a two-stage detection pipeline that integrates You Only Look Once Version 12 and MobileNetV2 to generate candidate bounding boxes, stabilized by a Kalman filter with a predict–update mechanism. This approach ensures robust and reliable object tracking, providing valuable insights into tennis training performance, placement accuracy, and actionable insights for sports analytics.

Keywords: computer vision; YOLO; object tracking; tennis training; sports analytics

1. Introduction

A tennis ball training system is designed to enhance training accuracy by reducing the dependence on subjective human visual judgment during practice in this study. Previous studies have indicated that accurately detecting ball impact points in racket sports is challenging, especially for small, fast-moving balls, which are prone to missed or erroneous detections [1]. In addition, traditional human observation is susceptible to factors such as attention lapses, fatigue, and subjective bias, often resulting in inconsistent training outcomes [2]. The developed system leverages deep learning techniques to enable real-time detection of tennis ball impacts on the wall, effectively overcoming challenges such as lighting variations, background noise, motion blur, and missed detections.

The automatic and reliable tennis ball tracking system operates under diverse environmental conditions, providing precise feedback on impact locations during wall-based drills. The system detects small, fast-moving tennis balls using You Only Look Once Version 12 (YOLOv12) [3] with a self-attention mechanism, enhancing recognition of small targets and improving temporal motion consistency across frames. Candidate bounding boxes are classified with the lightweight MobileNetV2, while an accuracy-based matching algorithm filters out background noise and false positives, improving robustness and real-time performance, especially in cluttered or outdoor environments. For dynamic tracking, a Kalman filter (KF) addresses jitter and occlusion by combining historical trajectories with current observations [4]. This enables smooth trajectory estimation and predictive

Check for updates

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 2 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

positioning, resulting in a stable and responsive tracking system suitable for real-time tennis training.

2. Related Work

2.1. YOLOv12

YOLOv12 is a YOLO-based object detection framework using attention mechanisms [5]. It includes the Area Attention Module (A2) for faster computation and the Residual Efficient Layer Aggregation Network for block-level residual connections, improving stability and convergence. Flash Attention optimizes memory, and unnecessary components are removed [6]. Similar YOLO frameworks are applied in sports, e.g., cricket, for shot detection and ball tracking [7].

2.2. MobileNetV2

MobileNetV2 is a lightweight CNN designed for efficient inference on resource-limited platforms. It has proven effective for object detection on low-power edge devices [8] and feature extraction for unmanned aerial vehicle localization [9]. In this study, MobileNetV2 is used to verify whether candidate bounding boxes correspond to actual tennis balls, ensuring robust recognition in real time.

2.3. KF

The KF is a recursive state estimation algorithm widely used in object tracking. Its core principle is to predict the next state based on the previous state and correct the prediction using the Kalman gain [10,11]. As shown in Equation (1), x'_t denotes the prior state estimate at time t, while $z_t - Hx'_t$ represents the innovation, i.e., the difference between the actual measurement z_t and the predicted measurement x'_t. The Kalman gain K_t determines the update magnitude, and multiplying it with the innovation yields the updated posterior state estimate $x_{\hat{t}}$

$$x_{\hat{t}} = x'_t + K_t\left(z_t - Hx'_t\right) \quad (1)$$

3. Methodology

3.1. Dataset Preparation

We prepared two datasets for training YOLOv12 and MobileNetV2, respectively. The overall process was divided into two stages: in the first stage, we constructed the training dataset for YOLOv12; in the second stage, under more complex environments, the bounding boxes detected by YOLOv12 were manually filtered to confirm whether they contained tennis balls. These filtered samples were then used to build both the training and testing datasets, which were subsequently fed into the MobileNetV2 model for further training.

3.1.1. Datasets for YOLOv12

The tennis dataset, from local practice and YouTube wall videos, consisted of 1902 annotated frames. Preprocessing included auto-orientation, resizing to 640 × 640, and contrast enhancement [12]. Data augmentation—horizontal flipping and ±15% brightness/exposure changes—tripled the dataset to 4492 images. The dataset was split into 70% training, 15% testing, and 15% validation. Figure 1 shows examples before and after preprocessing and augmentation.

3.1.2. Datasets for MobileNetV2

To perform binary classification (tennis ball vs. non-tennis ball), 500 images were collected from YOLOv12 candidate bounding boxes and categorized into Tennis Ball and Not Tennis Ball. The dataset was split 80/20 for training (400 images) and validation

(100 images), with a separate test set reserved. Images were resized to 224 × 224, converted to tensors, and normalized for MobileNetV2 input to improve generalization [13].



(**a**) (**b**)

Figure 1. Training images before and after preprocessing and augmentation: (**a**) Original Images; (**b**) Augmented and Standardized Images (resized to 640 × 640 pixels).

3.2. Ball Trajectory Tracking Algorithm

We developed a two-stage tennis detection and tracking method: YOLOv12 first detects candidate bounding boxes, then MobileNetV2 classifies them while the optimal tracking confidence selection algorithm filters true ball positions. These positions are input to a KF to estimate 3D position, velocity, and depth, enabling precise ball trajectory tracking [10,11], as shown in Algorithm 1.

Algorithm 1: Two-Stage Tennis Ball Detection and Tracking

Input: Video frames from a tennis wall practice
Output: Estimated ball positions($ball_x, ball_y, ball_z$), velocity($ball_v$)
Initialize KF with state vector: $\left(ball_x, ball_y, ball_z,\ ball_v\right)$
for each frame (time step Δt):
  Predict ball state using KF;
  Stage 1—YOLOv12 detection: Detect candidate bounding boxes $BB_i = (x_c, y_c, w, h)$;
  Stage 2—MobileNetV2 classification: validate candidates with probability $\geq T_{conf}$
  Select BB_i with highest *prob_ball* and tracking confidence;
  Update KF with selected BB_i position;
  if valid BB_i exist:
    $\Delta pos \leftarrow \left(ball_x^{(t)} - ball_x^{(t-1)}, ball_y^{(t)} - ball_y^{(t-1)},\ ball_z^{(t)} - ball_z^{(t-1)}\right)$;
    $ball_v = \frac{\Delta pos}{\Delta t}$;
    Return Estimated ball positions($ball_x, ball_y, ball_z$), velocity($ball_v$);
  else:
    Return previous state($ball_x, ball_y, ball_z$) and velocity($ball_v$);
  end if
end for

3.3. Ball-to-Wall Collision Detection

In wall practice scenarios, the system detects and tracks the tennis ball and estimates its motion states. At each frame, the position (x, y) and velocity components $\left(v_x, v_y\right)$ are

derived from the bounding box center (Cx, C_y) detected by YOLO and refined by MobileNetV2. These coordinates are fed into a KF with the state vector defined as follows.

$$State = (x, y, v_x, v_y, z, v_z) \tag{2}$$

A collision with the wall is determined when the vertical velocity v_y changes sign from negative to positive, indicating the ball's reversal after impact. To avoid multiple detections of the same bounce, a distance threshold d = 15 pixels is applied: a new rebound point (x_r, y_r) is only recorded if its distance from the last detected rebound point exceeds d. This ensures robust detection and accurate statistics of wall impact events. The mathematical condition can be summarized as follows.

$$(x_r, y_r) = \begin{cases} v_y(t-1) < 0 \; and \; v_y(t) > 0 \\ (x_r, y_r) \in WallZone \\ dist((x_r, y_r), (x_{last}, y_{last})) > d \end{cases} \tag{3}$$

4. Results and Discussions

4.1. Evaluation of Wall Impact Point Detection

To evaluate the system's accuracy in detecting tennis ball impact points on the wall, five one-minute wall practice videos were used as the test dataset. The dataset includes manually annotated ground-truth coordinates and system-detected coordinates, normalized using Euclidean distance [14]. Figure 2 shows the per-frame mean absolute error (MAE) between system-detected and ground-truth points, indicated by the red line. Most errors were below 35 pixels, with an overall MAE of 20.39 pixels, calculated as shown in Equation (4).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(X_{\text{detected},i} - X_{\text{gt},i}\right)^2 + \left(Y_{\text{detected},i} - Y_{\text{gt},i}\right)^2} \tag{4}$$

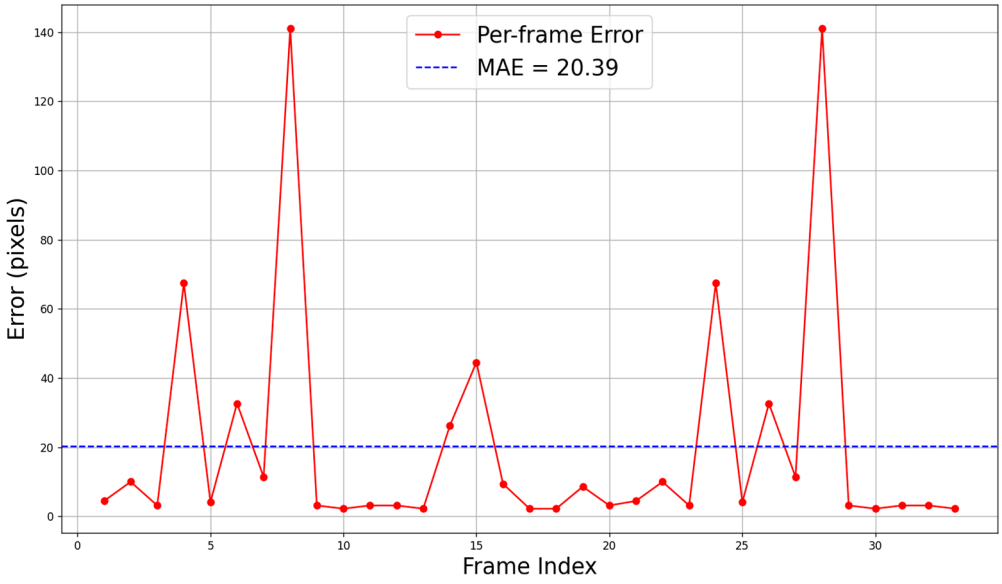


Figure 2. Per-frame error between system-detected and ground-truth impact points.

4.2. Model Evaluation

Using a test set of 3885 images, we evaluated both YOLOv12 and the two-stage MobileNetV2 approach. YOLOv12 achieves superior recall, successfully capturing the majority of tennis balls and maintaining stable detection performance even under challenging con-

ditions such as boundary proximity or varying illumination (Table 1). However, YOLOv12 exhibited a higher susceptibility to background noise, occasionally misclassifying non-ball objects such as fallen leaves on the court or boundary lines. Specifically, its false positives were 5.44 times higher compared to the two-stage filtering method (YOLO + MobileNetV2). In contrast, while the two-stage classifier reduced recall to 0.811, it achieved a markedly higher precision of 0.994, demonstrating its effectiveness in filtering out false positives and thereby improving the overall reliability of predictions. This trade-off highlights the design choice between maximizing detection coverage versus ensuring prediction reliability, depending on the target application scenario.

Table 1. Comparison of detection performance between YOLOv12 and two-stage models.

Model	True Positive	False Positive	False Negative	Precision	Recall	F1-Score
YOLOv12	3351	98	276	97.2%	92.4%	94.7%
YOLOv12 + MobileNetV2 (Two-Stage)	2774	18	647	99.4%	81.1%	89.3%

5. Conclusions and Future Works

We developed an AI-based model for tennis wall practice, capable of automatically detecting ball impact points and quantifying training performance. The two-stage filtering pipeline is adopted to reduce both false positives and false negatives, while integrating advanced detection models such as DEtection TRansformer and Shifted Window Transformer to improve the precision of candidate bounding boxes. Additionally, enhancements in illumination adaptation and frame interpolation will be explored to ensure robust system performance under dynamic and variable environmental conditions. Beyond tennis, the proposed framework shows promise for extension to other fast-paced ball sports and performance analytics scenarios, highlighting its potential for broader applicability across domains.

Author Contributions: Conceptualization, Y.-T.L., Y.-L.L., Y.-P.L. and S.-W.C.; methodology, Y.-T.L.; software, Y.-T.L. and Y.-P.L.; formal analysis, Y.-T.L. and Y.-P.L.; investigation, Y.-T.L. and Y.-L.L.; data curation, Y.-L.L. and S.-W.C. writing—original draft preparation, Y.-T.L. and Y.-L.L. writing—review and editing, P.-J.L. visualization, Y.-T.L.; supervision and funding acquisition, P.-J.L. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the National Science and Technology Council (NSTC) of Taiwan for financially supporting this research project under the grant NSTC114-2622-E-035-006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data may be available upon request.

Acknowledgments: The authors are grateful to the Hwa Kang Security Co., Ltd. (Taiwan) for provide experimental materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Naik, B.T.; Hashmi, M.F.; Bokde, N.D. A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions. *Appl. Sci.* **2022**, *12*, 4429. [CrossRef]
2. Öztürker, C.; Şahan, A.; Erman, K.A. Impact of Mental Fatigue on Tennis Players' Attention and Groundstroke Performance. *Front. Psychol.* **2025**, *16*, 1544785. [CrossRef] [PubMed]
3. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524.

4. Zhang, Y.; Xiao, Q.; Liu, X.; Wei, Y.; Xue, J. Method for Reconstructing Safety and Arming Motion Process by Integrating Kalman Filter and KCF. *Sci. Rep.* **2025**, *15*, 8334. [CrossRef] [PubMed]
5. Khanam, R.; Hussain, M. A review of YOLOv12: Attention-based Enhancements vs. Previous Versions. *arXiv* **2025**, arXiv:2504.11995. [CrossRef]
6. Ge, T.; Ning, B.; Xie, Y. YOLO-AFR: An Improved YOLOv12-based Model for Accurate and Real-time Dangerous Driving Behavior Detection. *Appl. Sci.* **2025**, *15*, 6090. [CrossRef]
7. Abbas, K.; Khan, M.I.; Saeed, M.; Ahmed, K.; Wang, H.; Mahmood, A. Front Pitch View Shot Extraction and Ball Tracking in Cricket using Deep Learning. *IEEE Access* **2025**, *13*, 110941–110954. [CrossRef]
8. Lokhande, H.; Ganorkar, S.R. Object Detection in Video Surveillance using MobileNetV2 on Resource-constrained Low-Power Edge Devices. *Bull. Electr. Eng. Inform.* **2025**, *14*, 357–365. [CrossRef]
9. Ooi, Y.M.; Chang, C.C.; Su, Y.M.; Chang, C.M. Vision-based UAV Localization on Various Viewpoints. *IEEE Access* **2025**, *13*, 38317–38324. [CrossRef]
10. Yang, J.; Jin, Y.J.; Kim, R.Y.C. Multi-Person Fall Detection Using Data Assimilation Method With Kalman Filter. *IEEE Access* **2025**, *13*, 104692–104705. [CrossRef]
11. Madhukar, P.S.; Madhukar, S. Kalman Filters in Different Biomedical Signals—An Overview. In Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 10–12 September 2020; pp. 1268–1272.
12. Deng, W.; Xie, G. Image Contrast Enhancement and Brightness Preservation based on An Adaptive Histogram Correction Framework. *Appl. Opt.* **2025**, *64*, 3502–3515. [CrossRef] [PubMed]
13. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
14. Wong, P.; Myint, H.; Hopgood, A. A Multi-View Automatic Table Tennis Umpiring Framework. *Proc. Inst. Mech. Eng. Part P J. Sports Eng. Technol.* **2023**, *238*, 339–351. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.