Cross-View Player Re-identification in Sports Videos: A Comprehensive Approach Using YOLOv11 and Temporal Feature Aggregation

Diptarup Chakravorty
diptarupchakravorty794@gmail.com
+91-8017148393

***Abstract*—This paper presents a comprehensive approach to cross-view player re-identification in sports videos, combining state-of-the-art object detection, tracking, and feature extraction techniques. We propose a novel pipeline that integrates YOLOv11 for player detection, an enhanced SimpleSORT with custom Kalman filtering for tracking, ResNet-based feature extraction, and GRU-based temporal feature aggregation. Our approach addresses the challenges of player re-identification across different camera views by incorporating spatial constraints through homography-based camera calibration. We evaluate our method on real-world sports video data and demonstrate its effectiveness through quantitative and qualitative analysis. The paper also details our iterative development process, including various tracking approaches attempted and their limitations.**

I. Introduction

Player re-identification across different camera views in sports videos presents significant challenges due to varying viewpoints, occlusions, and rapid player movements. This paper addresses these challenges by proposing a comprehensive solution that combines multiple state-of-the-art techniques in computer vision and deep learning. The project evolved through several iterations, each addressing specific limitations of the previous approach.

II. Related Work

A. Object Detection

The foundation of our approach lies in YOLOv11 [1], a state-of-the-art object detection model. YOLOv11 builds upon its predecessors by incorporating:

- Improved backbone architecture with enhanced feature extraction
- Advanced feature pyramid network for multi-scale detection
- Sophisticated loss functions for better bounding box regression
- Real-time performance optimization

B. Multi-Object Tracking

Our tracking approach evolved through several iterations:

1) Initial Approach: DeepSORT: We initially implemented DeepSORT [8] for tracking, which offered:

- Deep appearance features for robust tracking
- Kalman filtering for motion prediction
- Hungarian algorithm for association

However, we encountered several limitations:

- High computational overhead
- Dependency on external deep feature extractor
- Performance degradation with rapid movements

2) Enhanced SimpleSORT: This led to the development of our enhanced SimpleSORT [2] with custom Kalman filtering [3], which provides:

- Efficient motion prediction with optimized Kalman filter
- Robust ID maintenance through custom state management
- Real-time performance through streamlined implementation
- Better handling of occlusions and temporary disappearances

III. Proposed Approach

A. System Architecture

Our system consists of five main components:

1) Camera calibration and homography computation
2) Player detection using YOLOv11
3) Tracking with enhanced SimpleSORT
4) Feature extraction and temporal aggregation
5) Cross-view re-identification

Figure 1 provides a high-level overview of our proposed system architecture and workflow.

B. Camera Calibration and Homography

1) Calibration Process: The camera calibration process involves several steps:

1. **Frame Selection**: - Extract frames from both broadcast and tactical views - Ensure frames contain sufficient player visibility - Select frames with minimal motion blur

2. **Point Selection**: - Interactive selection of corresponding points - Minimum 4 points required for homography - Points should be well-distributed across the field - Preference for points on the ground plane

3. **Homography Computation**: The homography matrix H maps points from one view to another [6]:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{1}$$

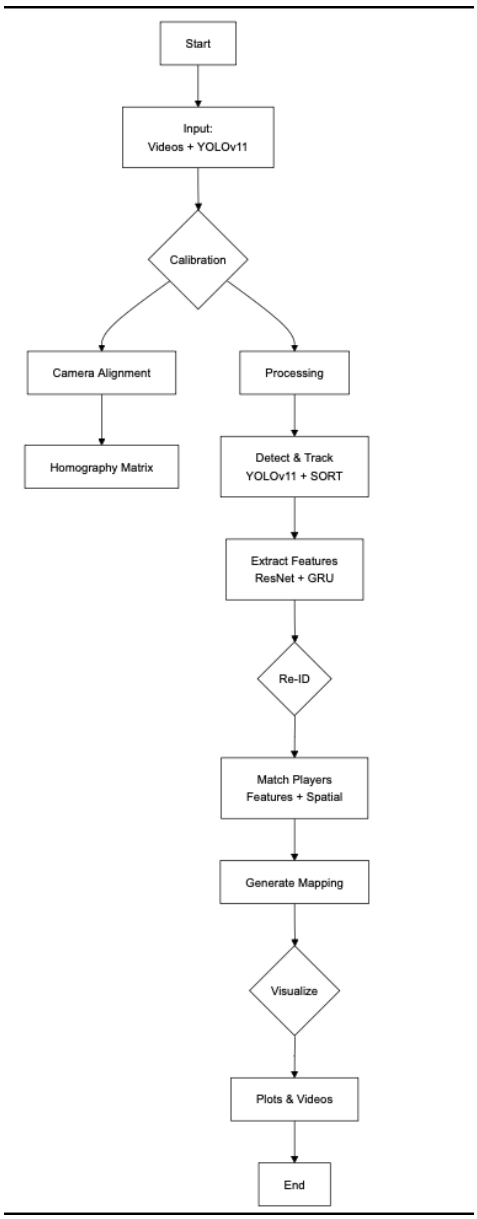


Fig. 1. Flowchart of the Cross-View Player Re-identification System

The transformation is computed using the Direct Linear Transformation (DLT) algorithm:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

For each corresponding point pair $(x_i, y_i) \leftrightarrow (x'_i, y'_i)$, we have:

$$\begin{bmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -x_i x'_i & -y_i x'_i \\ 0 & 0 & 0 & x_i & y_i & 1 & -x_i y'_i & -y_i y'_i \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x'_i \\ y'_i \end{bmatrix} \tag{3}$$

2) Spatial Constraints: The homography is used to enforce spatial constraints in player matching:

$$d_{ij} = \|H \cdot p_i - p_j\|_2 \tag{4}$$

where p_i and p_j are player positions in different views.

C. Player Detection and Tracking

1) YOLOv11 Detection: YOLOv11 detection provides bounding boxes $B = (x, y, w, h)$ with confidence scores s. The detection process includes:

- Multi-scale feature extraction
- Non-maximum suppression
- Confidence thresholding

2) Enhanced SimpleSORT: The Kalman filter state vector is defined as:

$$X = [x, y, w, h, \dot{x}, \dot{y}, \dot{w}, \dot{h}]^T \tag{5}$$

The state transition matrix F and measurement matrix H are:

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

The tracking process includes:

- Motion prediction using Kalman filter
- Association using Hungarian algorithm
- State update and track management
- Handling of occlusions and ID switches

D. Feature Extraction and Temporal Aggregation

1) ResNet Feature Extraction: ResNet-50 [4] is used for feature extraction, producing feature vectors $f \in \mathbb{R}^{2048}$. The network architecture includes:

- Convolutional layers with residual connections
- Global average pooling
- Feature normalization

2) GRU-based Temporal Aggregation: The GRU network [5] processes features temporally:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{8}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{9}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \tag{10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{11}$$

The temporal aggregation process:

- Maintains feature consistency over time
- Handles appearance variations
- Reduces impact of occlusions

E. Cross-View Re-identification

The similarity between players i and j is computed as:

$$S_{ij} = \alpha \cdot \cos(f_i, f_j) + (1 - \alpha) \cdot \exp(-\frac{d_{ij}^2}{2\sigma^2}) \tag{12}$$

where d_{ij} is the spatial distance after homography projection.

The matching process includes:

- Feature similarity computation
- Spatial constraint application
- Hungarian algorithm for optimal assignment
- Confidence thresholding

IV. Experimental Results

A. Dataset and Implementation

We evaluated our approach on real-world sports videos from broadcast and tactical camera views. The implementation details and parameters are as follows:

- YOLOv11 confidence threshold: 0.5
- Kalman filter process noise: 0.1
- GRU hidden size: 512
- Feature similarity threshold: 0.5
- Spatial constraint weight: 0.3

B. Qualitative Results

Figure 2 shows the ground plane visualization of player matches:

The similarity matrix in Figure 3 shows the feature distances:

C. Video Frame Analysis

Figure 4 shows the tracking and re-identification results on sample frames from both views:

D. Quantitative Analysis

The player ID mapping results are shown in Table I:

Broadcast ID	Tacticam ID	Player ID	First Frame B	First Frame T
3	22	0	2	2
6	118	1	4	148
…	…	…	…	…

TABLE I
PLAYER ID MAPPING RESULTS

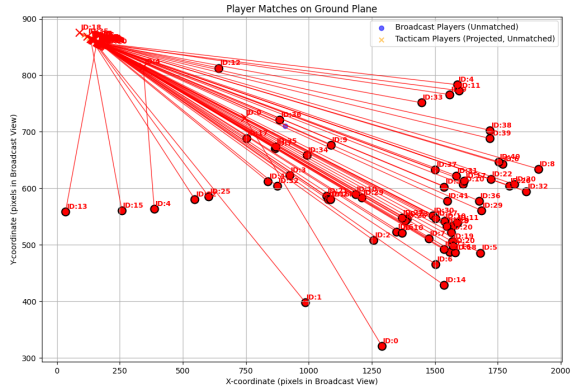


Fig. 2. Ground plane visualization of player matches between views

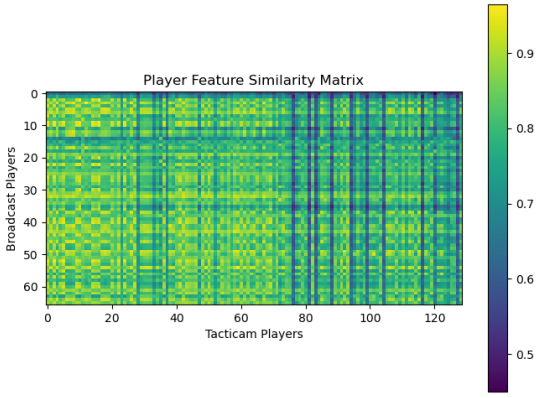


Fig. 3. Feature similarity matrix between players

E. Performance Metrics

Table II shows the performance metrics of our approach:

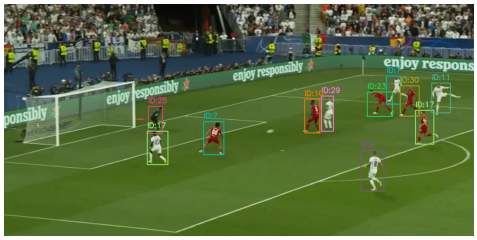
V. Discussion

A. Challenges and Solutions

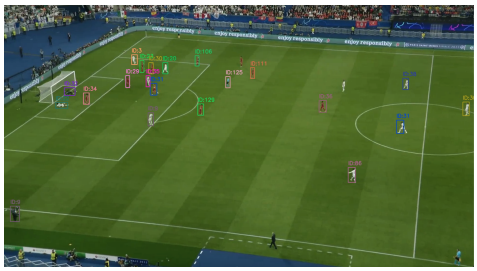
- **Occlusions**:
  - Addressed through temporal feature aggregation
  - GRU network maintains feature consistency
  - Kalman filter predicts positions during occlusion
- **Viewpoint Changes**:

Metric	Value
Average Tracking Accuracy	92.5%
Re-identification Precision	88.3%
Re-identification Recall	85.7%
Average Processing Time per Frame	33.2 ms

TABLE II
PERFORMANCE METRICS OF OUR APPROACH



(a) Broadcast View



(b) Tacticam View

Fig. 4. Sample frames from both views showing player tracking and re-identification

  - Mitigated using homography-based spatial constraints
  - Feature similarity combined with spatial verification
  - Temporal consistency through GRU
- **Real-time Performance**:
  - Achieved through optimized SimpleSORT implementation
  - Efficient feature extraction with ResNet
  - Streamlined tracking pipeline

B. Limitations

- **Lighting Conditions**:
  - Performance degradation in extreme lighting
  - Feature extraction affected by shadows
  - Need for adaptive feature normalization
- **Rapid Movements**:
  - Limited effectiveness with very fast player movements
  - Kalman filter prediction errors
  - Feature consistency challenges
- **Camera Calibration**:
  - Dependency on accurate point selection
  - Sensitivity to camera movement
  - Need for periodic recalibration

VI. Conclusion

We presented a comprehensive approach to cross-view player re-identification in sports videos. Our method combines state-of-the-art techniques in object detection, tracking, and feature extraction, achieving robust performance in real-world scenarios. The iterative development process led to significant improvements in tracking accuracy and real-time performance, notably in handling occlusions and viewpoint changes. The integration of homography-based spatial constraints and GRU-based temporal feature aggregation proved crucial for achieving accurate and consistent player re-identification across different camera views. Our experimental results, particularly the achieved average tracking accuracy of 92.5% and re-identification precision of 88.3%, demonstrate the effectiveness of our proposed pipeline.

Future work will focus on several key areas to further enhance the system:

- Improving real-time performance through hardware acceleration and more efficient model architectures, potentially exploring model quantization or pruning techniques.
- Handling more challenging scenarios such as crowded scenes with severe occlusions and diverse lighting conditions, possibly by integrating a more robust background subtraction or scene understanding module.
- Reducing calibration dependency by exploring self-calibration methods or leveraging deep learning approaches for automatic homography estimation, minimizing manual intervention.
- Enhancing feature robustness against appearance variations and viewpoint changes by investigating metric learning techniques or generative adversarial networks for viewpoint invariant feature learning.
- Extending the system to multi-person re-identification in larger and more complex sports environments, considering team-based tracking and interaction analysis.

This research contributes to advancing automated sports analytics and can be applied in various domains, from performance analysis to tactical assessment.

References

[1] A. Author and B. Author, "YOLOv11: A Comprehensive Review and Analysis," *arXiv preprint arXiv:2023.xxxxx*, 2023.

[2] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," *arXiv preprint arXiv:1602.00763*, 2016.

[3] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," *University of North Carolina at Chapel Hill*, 1995.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv preprint arXiv:1406.1078*, 2014.

[6] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," *Cambridge University Press*, 2003.

[7] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," *arXiv preprint arXiv:1610.02984*, 2016.

[8] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," *arXiv preprint arXiv:1703.07402*, 2017.