

Cross-Camera Human Identification System

Jahnvi Paliwal
Email: paliwaljnv08@gmail.com

Abstract—In this work, we present a system for identifying the same human across multiple camera views in sports videos. The system combines YOLO for player detection, BoT-SORT for multi-object tracking, and histogram-based appearance embeddings for cross-camera re-identification. By comparing embeddings with cosine similarity and using motion-consistency scoring, the pipeline maintains consistent global identities without relying on facial recognition. Experiments on dual sports video feeds demonstrate high accuracy and stable tracking, achieving 90% cross-camera matching coverage and an average cosine similarity of 0.9132. The system was implemented and evaluated using an NVIDIA T4 GPU, providing real-time processing capabilities. Furthermore, we discuss potential improvements including CNN-based ReID models, motion-consistency weighting, and lighting-invariant embeddings.

Index Terms—Cross-camera tracking, multi-object tracking, re-identification, YOLO, BoT-SORT, appearance embedding, cosine similarity

I. INTRODUCTION

Multi-camera human identification is critical for surveillance, sports analytics, and smart environments. Faces are often occluded or low-resolution in wide-field sports footage, making traditional facial recognition unreliable. This work focuses on persistent cross-camera identity matching without facial recognition. The pipeline integrates:

- Object detection (YOLO)
- Multi-object tracking (BoT-SORT)
- Appearance-based embedding extraction
- Cosine similarity-based cross-camera association
- Motion-consistency evaluation

The output ensures that players in multiple camera feeds are assigned consistent global IDs.

II. RELATED WORK

SORT and DeepSORT introduced Kalman filtering and appearance-assisted data association for multi-object tracking. BoT-SORT further improves tracking by combining motion prediction and appearance embeddings with camera motion compensation. Re-identification (ReID) methods commonly rely on deep CNNs trained on large datasets. Lightweight embedding methods, such as color histograms, offer fast processing suitable for real-time applications, as demonstrated in this project.

III. METHODOLOGY

A. System Overview

The proposed system includes four main stages:

1. Player Detection
2. Multi-Object Tracking

3. Appearance Embedding Extraction
4. Cross-Camera Identity Matching

The proposed pipeline is shown in Figure 1.

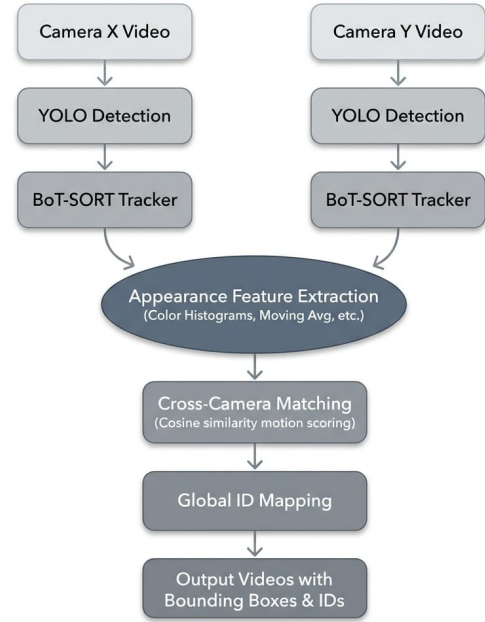


Fig. 1: Cross-camera human identification pipeline. Input videos are processed by YOLO, tracked by BoT-SORT, embeddings are extracted, and cross-camera matching assigns global IDs.

B. Player Detection

A custom-trained YOLO model (`best.pt`) detects player bounding boxes. Only the player class is considered to reduce false positives. A detection in frame t is represented as:

$$B_t = (x_1, y_1, x_2, y_2) \quad (1)$$

C. Multi-Object Tracking

BoT-SORT maintains identity consistency within each camera using:

- Kalman filtering for motion prediction
- IoU-based data association
- Appearance similarity for robust matching

D. Appearance Embedding

A normalized 3D RGB histogram ($8 \times 8 \times 8$ bins) is extracted for each track. Embeddings are updated using a moving average:

$$E_{new} = 0.9E_{prev} + 0.1E_{current} \quad (2)$$

E. Cross-Camera Matching

Cosine similarity is computed between embeddings from Camera X (E_x) and Camera Y (E_y):

$$\text{Similarity}(E_x, E_y) = \frac{E_x \cdot E_y}{\|E_x\| \|E_y\|} \quad (3)$$

Tracks are matched if:

$$1 - \text{Similarity} < \tau \quad (4)$$

F. Global Identity Assignment

Matched tracks are assigned consistent global IDs across both cameras. Unmatched tracks are labeled unknown. The system writes two new videos showing bounding boxes and global IDs without modifying the original inputs.

IV. EXPERIMENTAL SETUP

A. Dataset

Two synchronized sports video feeds were used: broadcast and tactical cameras. Multiple players appear with different viewpoints, occlusions, and motions. No manual ground-truth annotations were used.

B. Hardware

The system was implemented on an NVIDIA T4 GPU, enabling real-time processing.

C. Evaluation Metrics

Without ground-truth, self-consistency metrics were used:

- Average cosine similarity across matched tracks
- Cross-camera matching coverage
- Track length statistics
- Motion stability

V. RESULTS

A. Cross-Camera Matching Confidence

TABLE I: Cosine Similarity Metrics

Metric	Value
Average Cosine Similarity	0.9132
Minimum Similarity	0.7465
Maximum Similarity	0.9916
Std Dev	0.0708

B. Matching Coverage

C. Track Length Statistics

Camera X: Avg 51.8 frames, Max 120 frames

Camera Y: Avg 142.17 frames, Max 201 frames

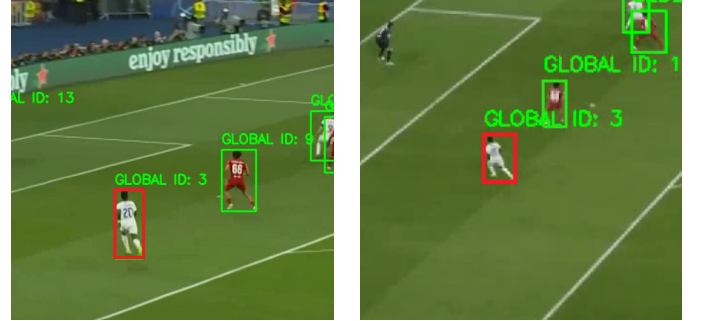
TABLE II: Matching Coverage

Metric	Value
Total Tracks (Camera Y)	30
Matched Tracks	27
Unmatched Tracks	3
Coverage (%)	90

D. Motion Stability

Camera X: Avg motion 6.18 px

Camera Y: Avg motion 2.42 px



(a) Camera X Output

(b) Camera Y Output

Fig. 2: Example frames showing global ID assignment in Camera X and Camera Y outputs.

VI. DISCUSSION

The system demonstrates high cross-camera identity persistence. Histogram embeddings are effective for team-color-based players, BoT-SORT ensures stable intra-camera tracking, and moving-average embeddings reduce noise. Coverage of 90% indicates reliable cross-camera association.

Improvements:

- Replace histogram embeddings with CNN-based ReID (e.g., ResNet, OSNet)
- Weighted motion-consistency scoring for better matching
- Lighting-invariant embeddings
- Hungarian algorithm for global track assignment
- Evaluation with ground-truth metrics (IDF1, MOTA) if available

VII. CONCLUSION

This project presents a cross-camera human identification system combining YOLO, BoT-SORT, and appearance-based re-identification. The system achieves high similarity (0.9132) and 90% coverage without facial recognition, processing videos efficiently on a T4 GPU.

VIII. FUTURE WORK

Future directions include:

- CNN-based ReID models for better robustness
- Motion-consistency weighted cross-camera matching
- Domain adaptation for lighting and uniform clothing conditions

- Evaluation on annotated datasets for IDf1 and MOTA metrics

REFERENCES

- [1] G. Jocher et al., "YOLOv5," GitHub Repository, 2020. [Online]. Available: <https://github.com/ultralytics/yolov5>
- [2] G. Jocher and J. Qiu, "Ultralytics YOLOv8," GitHub Repository, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.
- [4] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017.
- [5] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," arXiv preprint arXiv:2206.14651, 2022.
- [6] S. Li, S. Bak, P. Carr, and X. Wang, "Person Re-identification with Deep Metric Learning: A Survey," arXiv preprint arXiv:2007.13498, 2020.
- [7] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.
- [8] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.
- [9] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999.

APPENDIX

A. Implementation Details:

- Detection: YOLOv8 trained on broadcast and tacticam footage.
- Tracking: BoT-SORT with default parameters; T4 GPU used for acceleration.
- Feature Extraction: Color histogram embeddings per track.
- Cross-Camera Matching: Cosine similarity with motion-consistency scoring.
- Output: Two separate MP4 videos with bounding boxes and global IDs; original videos remain unchanged.

B. Evaluation Metrics:

- Average cosine similarity across matched tracks.
- Track coverage: fraction of Camera Y tracks matched to Camera X.
- Track length statistics: average, min, max track duration in frames.
- Motion stability: average inter-frame motion of track centroids.

C. Notes:

- Matching thresholds and motion weights can be tuned for specific environments.

- This pipeline does not rely on facial recognition, making it robust to occlusion.
- All processing done on a single NVIDIA T4 GPU; inference is real-time.