AI-Driven Soccer Analysis
Using Computer Vision

Turning Game Footage Into Actionable Coaching Data — Without Expensive Sensors

Adrian Manchado, Tanner Cellio, Jonathan Keane, Yiyang Wang
Milwaukee School of Engineering — Department of Computer Science

arXiv:2604.08722 — April 2026

What This Paper Is About

Professional soccer teams spend millions on GPS trackers, motion-capture systems, and dedicated analytics staff. But most amateur and college teams can't afford any of that — they might have nothing more than a camera recording games from the stands. This paper presents a system that takes ordinary game footage and automatically extracts useful statistics: where players are on the field, how fast they're moving, which team they're on, and how the team positions itself tactically. The entire system runs on AI — no special hardware required.

Why It Matters

Tactical insights like player heatmaps, sprint speeds, and formation analysis are typically only available to professional clubs with big budgets. This research democratizes access to those analytics. A college team with a single camera can now get data that was previously out of reach. It levels the playing field for resource-limited programs.

Key Results at a Glance

- **84.5% F1-score** for player detection — the system correctly identifies players in most frames
- **97.2% accuracy** for field keypoint detection — it knows exactly where it's looking on the field
- **0.5 meter average error** for mapping players to real-world field coordinates
- **17 out of 22 players** detected and tracked in a single game with zero manual fine-tuning
- **Works with any camera** — no GPS, no wearable sensors, just video footage

The Problem: Analytics Are Too Expensive

Modern soccer analytics rely on tracking systems that cost tens of thousands of dollars. GPS vests worn by players can track speed and position, but they require hardware on every player. Optical tracking systems use multiple cameras around the stadium — again, expensive. For a small college team or amateur club, these solutions are simply not an option.

What these teams DO have is a camera. Maybe it's mounted on a tripod, maybe a coach films from the stands. The question is: can we extract useful tactical data from just a video recording? This paper answers that question with a resounding yes.

What Kind of Data Can Be Extracted?

From a single video feed, the system can derive:

- **Player positions** — where every player is on the field at any moment
- **Team classification** — which team each player belongs to, determined automatically
- **Speed and distance** — how fast players move and how far they run
- **Positioning heatmaps** — visual maps showing where each team spends most of its time
- **Formation analysis** — the tactical shape each team adopts during play

How Is This Different From Previous Work?

Most existing computer vision sports systems are designed for professional leagues with labeled datasets and controlled camera setups. This system is built for the real world: raw, unprocessed footage with no labels, no preprocessing, and no guarantee about camera quality or angle. The researchers specifically wanted to show that useful analytics can come from messy, real-world data.

Feature	Professional Systems	This System
Cost	$50,000+ (multiple cameras)	Free (uses existing camera)
Hardware on players	GPS vests, IMU sensors	None required
Data preparation	Labeled, curated datasets	Raw, unprocessed footage
Camera requirements	Fixed, calibrated cameras	Any elevated camera
Setup time	Hours of calibration	Automatic — no setup needed
Team size	Full professional staff	One computer running AI

Table 1: How this system compares to professional tracking solutions

The AI Models Behind the System

The system combines four AI components, each handling a specific task. Think of them as specialists working together: one finds players, one tracks them, one knows the field, and one figures out which team is which.

1. Player Detection — Finding Players in Each Frame

The first challenge is identifying where players are in a video frame. The system tests four AI models for this task:

- **YOLOv5x** — A fast, single-pass detector. 'You Only Look Once' means it processes the entire image in one go, predicting bounding boxes around players instantly.
- **YOLOv8x** — An improved version with better accuracy for difficult cases like overlapping players.
- **YOLOv11x** — The newest version, using fewer computational resources while maintaining precision.
- **Faster R-CNN** — A two-stage detector that first proposes regions, then classifies them. More accurate but slower.

The key finding: YOLOv5x performed best for this application. It achieved the highest F1-score (0.845) — the most important metric because it balances catching all players (recall) without too many false alarms (precision). All three YOLO models outperformed Faster R-CNN, confirming that modern single-stage detectors are better suited for sports tracking.

2. Player Tracking — Following Players Across Frames

Detecting players in one frame is not enough — you need to track them as they move. This is where SAM2 (Segment Anything Model 2) comes in. SAM2 is Meta's powerful segmentation model that can identify and track any object across video frames.

The clever part: YOLO only needs to detect players in the FIRST frame. From that point on, SAM2 takes over, using its memory system to remember each player and track them throughout the game — even when players overlap, move behind each other, or briefly leave the camera view. This is much more efficient than traditional tracking methods that need to detect players in every single frame.

3. Field Key Point Detection — Knowing Where the Camera Is

To convert pixel positions into real-world distances, the system needs to understand the camera's perspective. It does this by detecting 12 key points on the soccer field — intersections of penalty arcs, center circles, and midfield lines.

A custom CNN (Convolutional Neural Network) was trained to find these points. The training data was manually labeled from 146 frames across two games with different weather conditions (glare and overcast). The model achieved 97.2% accuracy in determining which key points are visible and located them with an error of just 7.65 pixels — precise enough for reliable field mapping.

4. Team Classification — Who Plays for Which Team?

The simplest yet cleverest component: the system automatically figures out which team each player belongs to by analyzing the color of their jersey. For each detected player, the system samples the color at the center of their bounding box, then uses K-Means clustering (with k=2) to group players by color. No manual labeling needed.

Limitation: this approach can struggle with glare, shadows, or similar jersey colors. The researchers acknowledge this as an area for improvement.

How All the Pieces Fit Together

The complete system works like an assembly line. Each component feeds its output to the next, progressively transforming raw video into actionable tactical data.

The Workflow

Step 1: Input — Raw game footage from any elevated camera (tripod, stands, etc.)

Step 2: Player Detection — YOLOv5x identifies players in the first frame, generating bounding boxes

Step 3: Segmentation & Tracking — SAM2 uses YOLO's initial detections to track all players throughout the video

Step 4: Key Point Detection — A CNN identifies 12 reference points on the field in each frame

Step 5: Homography — Mathematical transformation converts camera-view coordinates to real-world field coordinates

Step 6: Team Classification — K-Means clustering assigns each player to a team based on jersey color

Step 7: Output — A 2D field map showing all players' real-world positions, separated by team

What Is Homography?

Homography is the mathematical bridge between what the camera sees and what's actually happening on the field. When a camera films from the stands, it creates a distorted perspective — players near the camera appear larger, and straight lines converge. Homography corrects this by calculating a transformation matrix that maps every pixel in the video to a real-world coordinate on the field.

The system uses the Direct Linear Transformation (DLT) algorithm, which solves a system of equations based on the 12 detected key points and their known positions on a standard NCAA soccer field (dimensions obtained from Google Maps). Once this matrix is computed, any player's position can be instantly converted from camera coordinates to meters on the field.

Pipeline Stage	Input	Output	Model Used
Player Detection	Raw video frame	Bounding boxes around players	YOLOv5x
Segmentation & Tracking	Bounding boxes	Pixel-accurate player masks with IDs	SAM2
Key Point Detection	Raw video frame	12 field reference points	Custom CNN
Homography	Key points + field template	Transformation matrix	DLT algorithm
Team Classification	Player bounding box colors	Team assignments (2 groups)	K-Means

Table 2: The complete pipeline from video to tactical data

Results: How Well Does It Work?

Player Detection Performance

Four models were evaluated on 390 annotated frames from the MSOE men's soccer team's 2024 season. The ground truth was created using SAM2 to manually segment players, then converting those masks into bounding boxes.

Model	Mean IoU	Recall	Precision	F1-Score
Faster R-CNN	0.6837	0.6574	0.7942	0.7194
YOLOv11x	0.7934	0.6805	0.9229	0.7834
YOLOv8x	0.7466	0.8152	0.8765	0.8447
YOLOv5x (winner)	0.7644	0.7995	0.8963	0.8451

Table 3: Object detection model comparison — YOLOv5x wins overall

YOLOv5x achieved the best overall balance. It correctly identified 17 out of 22 players without any fine-tuning. The remaining 5 players were missed due to distance, overlap, or poor lighting. All three YOLO variants significantly outperformed the older Faster R-CNN architecture.

Key Point Detection Performance

Metric	Training Set	Testing Set
Visibility Accuracy	99.89%	97.18%
MAE (normalized)	0.0107	0.0138
MAE (pixels)	5.96	7.65

Table 4: The CNN correctly identifies and locates field key points

The model correctly determines whether a key point is visible in 97.2% of cases. When it locates a point, it's off by an average of just 7.65 pixels — very precise for this application. The small gap between training and testing performance shows the model generalizes well and isn't overfitting.

Full System Accuracy

When the entire pipeline is combined (detection + tracking + key points + homography), the system maps players to real-world coordinates with an average error of:

- **0.225 meters** for ground truth key points (the reference points themselves)
- **0.26 meters** for predicted key points (what the CNN detects)
- **0.499 meters** overall projection error for transformed player positions

An error of about half a meter is quite good for a system running on a single camera without any calibration hardware. For context, a soccer player's body width is about 0.5 meters, so the system can roughly distinguish individual player positions.

The Data: What Was Used to Build and Test the System

Game Footage

The dataset consists of footage from 10 home games of the MSOE (Milwaukee School of Engineering) men's soccer team during the 2024 season. The footage was recorded using a high-quality BePro camera mounted at an elevated position, providing good field coverage. Games were played under diverse conditions: daytime, nighttime, and rainy weather. The camera frequently zoomed, panned, and shifted to follow the ball, creating the kind of dynamic, challenging footage that real-world systems must handle.

How Ground Truth Was Created

To evaluate the AI models, the researchers needed 'ground truth' — correct answers to compare against. This is where things get interesting, because creating labeled data for sports is notoriously time-consuming.

For player detection, they used SAM2 itself as a labeling tool. By providing SAM2 with initial prompt points (generated by YOLO), they could quickly segment players in 390 consecutive frames. Each segmentation mask was then converted into a bounding box. This semi-automated approach significantly reduced the manual effort required.

For field key points, there was no shortcut — the researchers manually labeled 12 reference points across 146 frames from two different games (one with glare, one overcast). Each frame typically had 4–12 visible key points depending on the camera angle. This small but carefully curated dataset was sufficient to train an effective keypoint detection model.

Data Augmentation

With only 146 labeled frames for keypoint detection, the dataset was small. To make the most of it, the researchers applied horizontal flipping — mirroring each image and adjusting the keypoint coordinates accordingly. This effectively doubled the dataset while teaching the model to recognize field patterns regardless of left-right orientation. The final dataset was split 80/20 for training and testing.

Training Details

The keypoint CNN was trained using the Adam optimizer with an exponentially decaying learning rate. A Bayesian optimization search was performed to find the best hyperparameters — dropout rate, layer sizes, and learning rate. The custom loss function calculated error only for visible key points (ignoring those outside the frame), ensuring that the model wasn't penalized for correctly identifying that a point isn't visible.

Understanding the Technology

To fully appreciate what this system does, it helps to understand the underlying technology. This section explains the key concepts in plain language.

How Does Computer Vision Work?

Computer vision is the field of AI that gives computers the ability to understand images and videos — just like humans can look at a photo and identify people, objects, and actions. The core idea is to train neural networks on thousands of labeled images so they learn to recognize patterns.

A neural network processes images through layers. The early layers detect simple features like edges and colors. Middle layers combine these into more complex patterns like shapes and textures. The final layers recognize objects — 'this is a person,' 'this is a soccer ball,' 'this is a field marking.' The network learns these patterns by adjusting its internal parameters during training, gradually improving its accuracy.

Why YOLO Is Fast

Traditional object detection methods work in two stages: first they propose regions that might contain objects, then they classify each region. This is like searching a room by checking every corner one by one.

YOLO (You Only Look Once) takes a fundamentally different approach. It divides the image into a grid and predicts bounding boxes and class probabilities for every grid cell simultaneously. This single-pass approach makes YOLO dramatically faster — it can process 30–60 frames per second, which is fast enough for real-time video analysis.

What Makes SAM2 Special

SAM2 (Segment Anything Model 2) is Meta's breakthrough in video segmentation. Unlike earlier models that processed each frame independently, SAM2 has a 'memory' that remembers what it saw in previous frames. This is crucial for tracking because it allows the model to maintain a consistent identity for each player even when they move, get partially hidden, or change appearance due to lighting.

The 'segmentation' part means SAM2 doesn't just draw a box around a player — it identifies the exact pixels that belong to each player, creating a precise mask. This pixel-level accuracy enables much more precise position measurements than simple bounding boxes.

How Homography Corrects Perspective

When you photograph a rectangular soccer field from the stands, the result is a trapezoid — the far goal appears smaller than the near one, parallel lines seem to converge, and distances are distorted. Homography is the mathematics that undoes this distortion.

By identifying known points on the field (like where the penalty arc meets the sideline), the system calculates a transformation matrix. This matrix can then convert ANY point in the camera image to its true position on the field. It's like having a magic pair of glasses that shows the field from directly above, regardless of where the camera is actually positioned.

Understanding the Evaluation Metrics

The paper uses several metrics to measure performance. Here's what each one means:

- **Precision**: Of all the players the AI detected, what percentage were actually players? High precision means few false alarms.
- **Recall**: Of all the real players in the frame, what percentage did the AI find? High recall means few missed players.

F1-Score: The balance between precision and recall. The single best metric for overall detection quality.

IoU (Intersection over Union)): How well the predicted bounding box overlaps with the actual player position. A score of 1.0 means perfect alignment.

MAE (Mean Absolute Error)): The average distance between predicted and actual positions. Lower is better.

What Can Coaches Do With This Data?

Once the system maps player positions to real-world coordinates, a wealth of tactical insights become available. Here's what coaches can analyze:

Speed and Distance Metrics

By tracking a player's position across consecutive frames (at known time intervals), the system calculates instantaneous speed and total distance covered. This is the same type of data that GPS vests provide — but derived entirely from video.

Positioning Heatmaps

By accumulating a player's positions over time, the system generates heatmaps showing where each player spends most of the game. Coaches can see if a winger is staying too wide, if a defender is pushing too far forward, or if the team is leaving gaps in the middle of the field.

Formation Analysis

By averaging positions of all players on a team, the system can identify the tactical formation (4-4-2, 4-3-3, etc.) and track how it shifts during different phases of play. Does the team press high or sit deep? Do they maintain their shape when defending? These questions can be answered with data.

Team-Level Statistics

Beyond individual players, the system provides team-wide metrics: average defensive line height, team compactness (how close players stay to each other), and territorial dominance (which areas of the field each team controls).

Metric	What It Shows	Coaching Value
Player Speed	How fast each player runs	Fitness assessment, sprint training targets
Distance Covered	Total kilometers per player	Workload management, substitution timing
Heatmaps	Where players spend time	Positioning adjustments, tactical gaps
Formation Shape	Team tactical structure	Strategic planning, opponent analysis
Defensive Line Height	How far back/forward the defense sits	Pressing strategy, offside trap planning
Territorial Control	Which areas each team dominates	Possession strategy, field positioning

Table 5: Tactical insights available from the system

Current Limitations

The researchers are honest about what doesn't work perfectly yet. Understanding these limitations is important for anyone thinking about using this system.

False Positives

The detection model sometimes identifies non-players as players — ball boys, referees, or even spectators near the field. While SAM2's size filtering removes most false positives, some still slip through.

Team Classification Errors

The jersey-color clustering approach struggles when lighting changes dramatically (glare, shadows) or when teams have similar-colored jerseys. Players can be misclassified when they move between well-lit and shaded areas.

Player Re-Identification

When a player leaves the camera frame and returns, the system doesn't automatically recognize them as the same person. Each entry creates a new tracking ID, which makes it impossible to calculate individual player statistics across the entire game.

Field Generalization

The keypoint detection model was trained only on footage from MSOE's home field. It may not generalize well to other fields with different camera angles, backgrounds, or markings. This is a significant limitation for away games.

No Ball Tracking

The current system tracks players but not the ball. Without ball tracking, important statistics like passes, shots, and possession chains cannot be calculated.

Single Camera Limitation

Using one camera means players can be occluded (hidden behind other players) from certain angles. Professional systems use multiple cameras to eliminate blind spots.

Future Directions

The researchers have a clear roadmap for improving the system. Each improvement addresses a specific limitation while building on the existing foundation.

Player Re-Identification

When a player exits and re-enters the frame, the system currently treats them as a new person. The solution is to maintain a memory of each player's last known position, team assignment, and visual appearance. When a new player appears near where an old one left, the system will check if they match — using color analysis and movement trajectory. If they match, the original tracking ID is restored, enabling continuous individual statistics throughout the game.

Multi-Field Training

The keypoint detection model was trained exclusively on MSOE's home field. While it performs well there, it may struggle at away games with different camera angles, field markings, or backgrounds. The solution is straightforward: collect and label footage from multiple fields under various conditions. This will teach the model to recognize soccer field features regardless of the specific venue.

Ball Detection and Tracking

Adding ball detection is the highest priority for expanding the system's capabilities. With ball tracking, the system could calculate passes (ball moving between teammates), shots (ball moving toward goal), and possession chains (sequences of passes). These are the most tactically valuable metrics in soccer and would make the system significantly more useful for coaches.

Player-Level Statistics

Once re-identification is solved, individual player statistics become possible: total distance covered, number of sprints, time spent in each field zone, and contribution to team formation. This data is invaluable for player development, fitness management, and selection decisions.

Real-Time Processing

Currently the system processes recorded footage after the game. The researchers aim to optimize the pipeline for real-time operation, enabling live tactical feedback. A coach could watch a tablet showing a 2D field map with live player positions, receiving immediate insights about formation shifts, pressing intensity, or defensive gaps.

Integration With Existing Tools

The system could be integrated with existing coaching software, video analysis tools, and team management platforms. By providing a standard API for the extracted data, the system becomes a plugin that enhances rather than replaces current workflows.

Broader Impact and Applications

While this paper focuses on soccer, the underlying approach is sport-agnostic. The same pipeline — detect objects, track them, map to real-world coordinates — could be applied to any sport filmed from a single camera.

Other Sports

- **Basketball** — Track player movement, analyze pick-and-roll efficiency, measure spacing
- **Field hockey** — Similar field dynamics to soccer, same analysis applies
- **Rugby** — Track formation patterns, analyze defensive line speed
- **Cricket** — Track fielding positions, measure bowling run-ups

Beyond Sports

The technical approach — combining object detection, tracking, key point detection, and homography — has applications in surveillance, autonomous vehicles, robotics, and augmented reality. Any domain that needs to convert camera-view observations into real-world coordinates can benefit from this methodology.

For example, traffic monitoring systems could use similar techniques to track vehicles and calculate speeds from standard traffic cameras. Warehouse management could track workers and equipment for safety and efficiency. Wildlife researchers could monitor animal movements from drone footage. The core pipeline — detect, track, map to real coordinates — is universally applicable.

The Research Team's Motivation

This project originated from a real need. The MSOE men's soccer team wanted tactical analytics but couldn't afford professional tracking systems. Computer science students teamed up with the soccer program to build a solution. This collaboration between technologists and athletes is a model for how university programs can leverage student talent to solve real-world problems while providing valuable learning experiences.

Open Science

The paper is published on arXiv as an open-access preprint, meaning anyone can read, use, and build upon this work. The approach is designed to be reproducible — other teams can implement the same pipeline with their own footage. This openness accelerates progress in the field and makes the technology accessible to everyone.

Democratizing Sports Analytics

The most significant impact of this work is making analytics accessible. Currently, only top-tier professional clubs can afford GPS tracking, multiple camera setups, and dedicated analytics teams. This system could put meaningful tactical data in the hands of college teams, youth academies, and amateur clubs — anyone with a camera and a computer.

Industry	Current Cost of Tracking	Potential With This System
Professional Soccer	$100K+ (GPS + multi-cam)	Supplementary analysis from broadcast footage
College Soccer	$0–$5K (basic cameras only)	Full tactical analytics for the first time
Youth Academies	$0	Player development tracking from parent-filmed games

Amateur Clubs	$0	Post-game analysis from smartphone recordings
Sports Broadcasting	Existing multi-cam setups	Automated real-time statistics overlay

Table 6: Cost reduction potential across different levels of play

Conclusion

This paper demonstrates that meaningful soccer analytics can be extracted from ordinary game footage using AI. By combining YOLO for player detection, SAM2 for tracking, a custom CNN for field recognition, and homography for coordinate mapping, the system transforms a single camera feed into a tactical analysis tool.

The key achievement is that this works on raw, unlabeled footage — no expensive hardware, no pre-labeled datasets, no professional camera operators. While limitations remain (player re-identification, field generalization, ball tracking), the foundation is solid and the roadmap for improvement is clear.

For resource-limited teams, this represents a paradigm shift: professional-level analytics without the professional-level budget.

Glossary: Key Terms Explained

YOLO (You Only Look Once): An AI model that detects objects in images by looking at the entire image once and predicting all bounding boxes simultaneously. Fast enough for real-time use.

SAM2 (Segment Anything Model 2): Meta's AI model that can identify and track any object across video frames with pixel-level precision. Uses a memory system to remember objects over time.

Faster R-CNN: A two-stage object detection model that first proposes regions of interest, then classifies what's in each region. More accurate but slower than YOLO.

CNN (Convolutional Neural Network): A type of AI specifically designed for analyzing images. It learns to recognize patterns by processing visual data through layers of filters.

Homography: A mathematical transformation that maps points from one 2D plane (camera view) to another (real-world field). Corrects perspective distortion.

F1-Score: A single number that balances precision (how many detections are correct) and recall (how many real objects were found). Higher is better.

IoU (Intersection over Union): Measures how well a predicted bounding box overlaps with the actual object. A perfect match scores 1.0.

K-Means Clustering: An algorithm that groups data points into K clusters based on similarity. Here, it groups players by jersey color into 2 teams.

Bounding Box: A rectangle drawn around a detected object in an image, marking its position and size.

Segmentation: Identifying exactly which pixels in an image belong to an object (not just drawing a box around it).

MAE (Mean Absolute Error): The average distance between predicted and actual positions. Lower means more accurate.

DLT (Direct Linear Transformation): An algorithm for computing homography by solving a system of linear equations from corresponding point pairs.

Original Paper Reference

Manchado, A., Cellio, T., Keane, J., & Wang, Y. (2026). AI Driven Soccer Analysis Using Computer Vision. arXiv:2604.08722. https://arxiv.org/abs/2604.08722

Simplified version created for general audience accessibility.