A Survey on Applications of Deep Learning Algorithms in Basketball and Soccer

Daksh Sangal
IIIT Kota
2023kucp1010@iiitkota.ac.in

Ankit Goyal
IIIT Kota
2023kucp1003@iiitkota.ac.in

Mahesh Waghmare
IIIT Kota
2023kucp1004@iiitkota.ac.in

Abstract

Deep learning and sports analytics represent a rapidly developing field among researchers. However, limited surveys exist in this domain and even fewer focus specifically on sport oriented research applications. To address this gap, we have combined two similar and highly popular sports, basketball and football, as our primary domains of investigation. We are trying to compare different researches in the area of application of deep learning in sports, examining how various neural network architectures and machine learning techniques have been employed to solve complex problems in game analysis and prediction. Our goal would be to compare research on three main topics for now: first, 2D to 3D modeling of games, which involves reconstructing three-dimensional spatial information from two dimensional video footage; second, Game Action Categorization, which focuses on identifying and classifying specific events, movements, and tactical maneuvers during game play; and third, Game Actions predictions or winning predictions, which encompasses forecasting future plays, player decisions, and overall match outcomes. We will compare scores of different models used in different researches across these three application areas, presenting the performance metrics, methodologies, and results in a comprehensive manner, and let the reader decide which is best suited for their specific use case or research direction.

1. Introduction

Sports analytics powered by deep learning represents an area where significant research has been conducted to improve the understanding and analysis of athletic performance. However, despite these efforts, this remains a field with limited comprehensive surveys and is still very much a developing domain. There is considerable scope for advancement in this area, and we aim to help with this survey by comparing different research that has happened since the early applications of deep learning in sports and examining what the future prospects might be for researchers looking to contribute to this field.

For our analysis, we have chosen soccer and basketball as the primary sports to focus on because they are among the most popular sports worldwide. These sports are kind of similar in nature, both are ball oriented team sports with dynamic gameplay, and importantly, much more research has been conducted in these two sports due to their global popularity and high consumer demand for advanced analytics. Therefore, we hope that conducting a survey that focuses primarily on these sports but is not limited

to them alone would be able to generate the greatest impact and provide valuable insight that could potentially be extended to other sports as well.

After studying numerous researches in this domain, we have identified some key areas in sports analytics that researchers are actively trying to solve with deep learning techniques. Some of the major areas we will focus on are: first, 2D to 3D conversion, which involves transforming flat video footage into three dimensional spatial representations; second, Game Prediction and Win Prediction, which uses historical and real time data to forecast match outcomes; and third, Game Action Categorization, which involves identifying and classifying specific events and movements during gameplay.

What our aim is to systematically summarize the researches that different people and research groups have done on these topics. We will list their goods and bads, highlighting the strengths and limitations of each approach detail the models they used, examine the different approaches they took to solve similar problems, and document the datasets they used for training and validation. We will provide this comparison to the reader in a nice tabular and visual way, making it easy to understand and compare multiple studies at a glance. We hope to help fellow researchers with these comparisons so that if in the future someone extends this research or begins new work in these areas, they can take the best path forward by learning from previous successes and avoiding known pitfalls.
Additionally, we will be linking references to all the study material we read and analyzed, providing access to the datasets that were used by the researches that we mention in our survey, and including clear definitions for technical keywords and terminology to ensure our survey is accessible to both newcomers and experienced researchers in the field.

2. KeyWords

This section will act as a reference to the reader providing quick definitions and explanations for terms commonly used in deep learning. this section is indented for readers that are not well versed in the field of deep learning and the explanations will reflect our intent. All the keywords whose definitions are mentioned here will be written in italics.

- ReLu
- Sigmoid
- Negative Log Loss
- ECE
- Spatio Temporal Data
- LSTM
- tanh

3. Researches

This section forms the core of the survey and provides a comprehensive examination of the research studies selected for analysis. To ensure clarity and meaningful comparison, the works are organized into three major thematic categories that reflect distinct methodological goals within sports analytics research: Game Prediction, Game Action Categorization, and 3D Game Simulation. Each thematic category captures a different dimension of how deep learning and computer vision have been applied to soccer and basketball, ranging from forecasting micro-actions and possession outcomes, to identifying complex in-game events, to reconstructing 3D representations of players and gameplay from 2D video.

Within each category, the studies are further subdivided into focused subsections, where each subsection is devoted to a single research paper. This structure enables readers to explore the nuances of each method while maintaining a coherent progression across the broader research landscape.

For every study included in the survey, we present a systematic and critical summary based on five key components:

1. a clear articulation of the problem being addressed and the authors' research objectives;
2. an explanation of the methodological framework and theoretical foundations;
3. a detailed description of the datasets used, including how they were collected, annotated, and prepared;
4. an overview of the machine learning models or algorithms implemented; and

) a summary of the empirical results, evaluation metrics, and conclusions derived from the findings.

This consistent analytical format provides a unified lens through which to evaluate diverse methodologies, making it easier to compare approaches, understand their contributions, and identify overarching trends across the literature. Additionally, it highlights the strengths and limitations of each study, supporting a holistic understanding of how deep learning is shaping modern sports analytics.

3.1. Game Prediction

This category highlights research that aims to predict the probabilities of micro-actions occurring within a game, using detailed spatio-temporal information such as player positions, ball trajectories, team formations, and lineup identities. Micro-actions—such as passes, dribbles, defensive rotations, screens, or subtle positional adjustments—are fundamental components of both soccer and basketball, yet traditionally remain unquantified in standard analytics. Recent deep learning approaches attempt to model these actions probabilistically by learning from large-scale tracking and event datasets, enabling analysts to estimate expected outcomes, assess decision-making, and understand tactical structures at a granular level. This section presents two representative studies, one focusing on soccer and the other on basketball, each of which develops a deep predictive framework to estimate the likelihood and value of micro-actions based on evolving game context. For both works, we summarize the problem formulations, methodological frameworks, datasets used, model architectures, and key evaluation metrics. Together, these studies demonstrate how modern deep learning pipelines—particularly those combining convolutional, recurrent, and contextual modeling—provide powerful tools for forecasting in-game behavior and quantifying micro-level contributions that were previously invisible in traditional sports analysis.

3.1.1. SoccerMap [[1]]

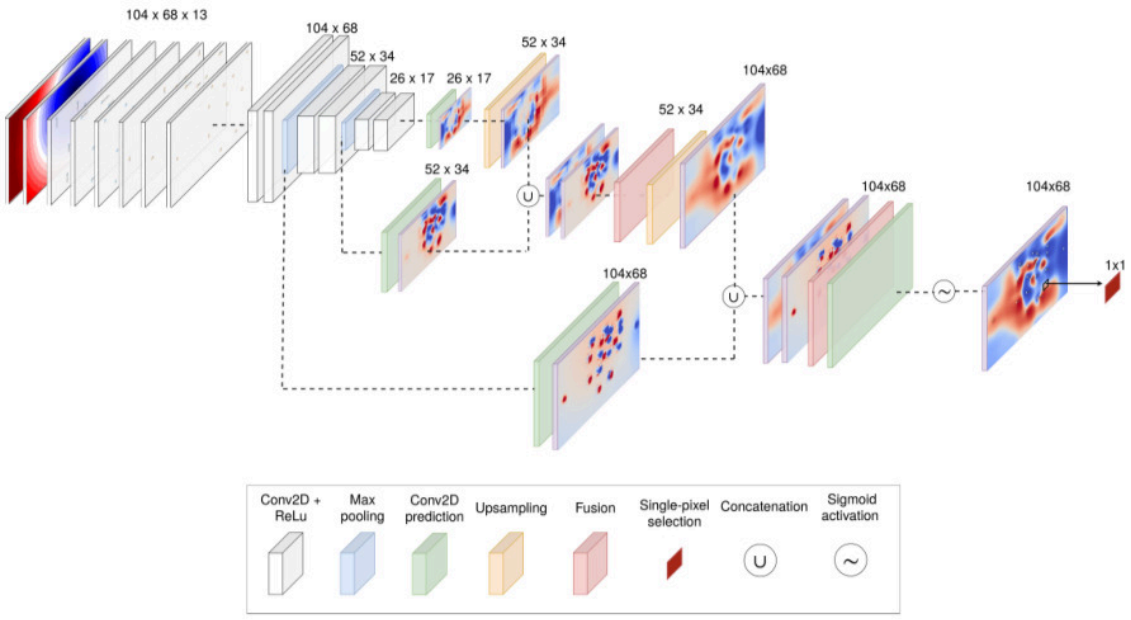

Figure 1: Architecture of the SoccerMap model.

3.1.1.1. Objective

Fernández and Bornn (2020) [[1]] study the problem of estimating a full probability surface over the soccer pitch that represents, for every possible target location, the likelihood that a pass originating from the current game state will be completed successfully. Traditional soccer analytics tend to focus on discrete, realized events—such as the passes that were actually made, shots taken, or possession transitions—and therefore overlook the rich space of potential decisions available to a player. The authors argue that coaches and analysts require a model that evaluates not only what happened but what could have happened, motivating a dense prediction approach based on fully convolutional networks [[2]]. Because each training instance contains only a single positive pixel, the task requires learning from extremely sparse labels, making it an instance of weakly supervised dense prediction [[3]]. SoccerMap is therefore proposed as a high-resolution, spatially coherent model capable of producing interpretable probability maps.

3.1.1.2. Methodological Framework

The authors represent each game state as a tensor of shape l×h×c, where each channel encodes low-level features derived from tracking data [[1]]. The architecture is fully convolutional in the sense of Long et al. [[2]], producing a dense output for every grid cell without collapsing spatial structure. SoccerMap processes the input at three spatial resolutions (1x, 1/2x, and 1/4x), reflecting modern multi-scale feature extraction strategies common in residual networks [[4]]. At each scale, two symmetric-padded 5×5 convolution layers followed by ReLU activations [[5]] capture local spatial patterns while max-pooling layers expand the receptive field.

Lower-resolution predictions are upsampled using learned, nonlinear modules and fused with high-resolution predictions in a multi-scale integration stage [[1]]. Final probabilities are generated through

a 1×1 convolution and sigmoid activation. Training applies a log-loss objective at only the ground-truth destination pixel, an approach aligned with weakly supervised dense labeling frameworks [[3]].

3.1.1.3. Dataset

The dataset consists of high-frequency player and ball tracking data sampled at 10 Hz, aligned with manually annotated pass events [[1]]. For each pass, the frame corresponding to ball release is labeled with a single grid-cell target, while all other cells remain unlabeled. This sparsity is the core motivation for the weakly supervised formulation.

Data spans 740 English Premier League matches (2013/14–2014/15) provided by STATS LLC [[1]]. The pitch is discretized to approximately 104×68, with 13 feature channels encoding distances, angles, pressure metrics, and spatial relationships between players. This produces a large and diverse corpus enabling dense learning across thousands of unique pass contexts.

3.1.1.4. Models Implemented

The primary model is the SoccerMap architecture [[1]], which integrates ideas from fully convolutional networks [[2]], multi-scale feature hierarchies [[4]], and dense prediction fusion strategies. Symmetric padding ensures geometry-preserving convolutions, preventing edge distortions. Training employs log-loss at the target pixel, relying on spatial coherence to propagate meaningful gradients across the entire prediction map.

Baselines include Logistic Net and Dense2 Net [[1]], both of which collapse the spatial dimension and produce a single scalar prediction. Ablation studies confirm that removing multi-scale processing significantly reduces predictive quality.

3.1.1.5. Results and Conclusions

The dataset is divided into 60% training, 20% validation, and 20% test splits [[1]]. Evaluation uses negative log-loss and Expected Calibration Error (ECE). SoccerMap achieves a log-loss of 0.217 and ECE of 0.0225, surpassing both traditional baselines.

Most importantly, the generated probability surfaces provide interpretable insights into decision-making, highlighting optimal lanes, risky zones, and defensive pressure patterns. The authors conclude that SoccerMap provides both superior predictive performance and valuable analytic interpretability, and suggest future extensions involving richer contextual features and expanded dense labeling strategies [[1]].

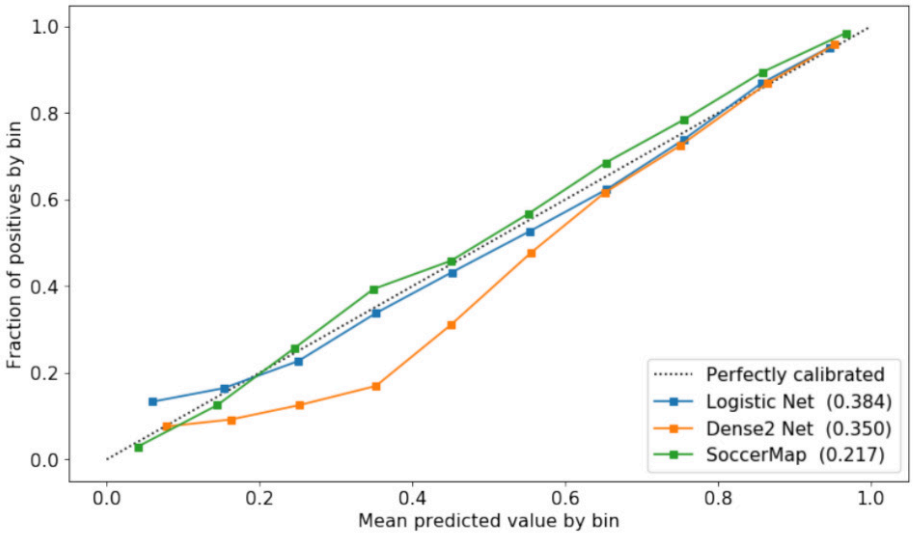

Figure 2: Visual comparison of model outputs on pass-probability surfaces.

Model	Log-loss	ECE	Inference time	Number of parameters
Naive	0.5451	-	-	0
Logistic Net	0.384	0.0210	0.00199s	11
Dense2 Net	0.349	0.0640	0.00231s	231
Soccer Map	0.217	0.0225	0.00457s	401, 259

3.1.2. DeepHoops [[6]]

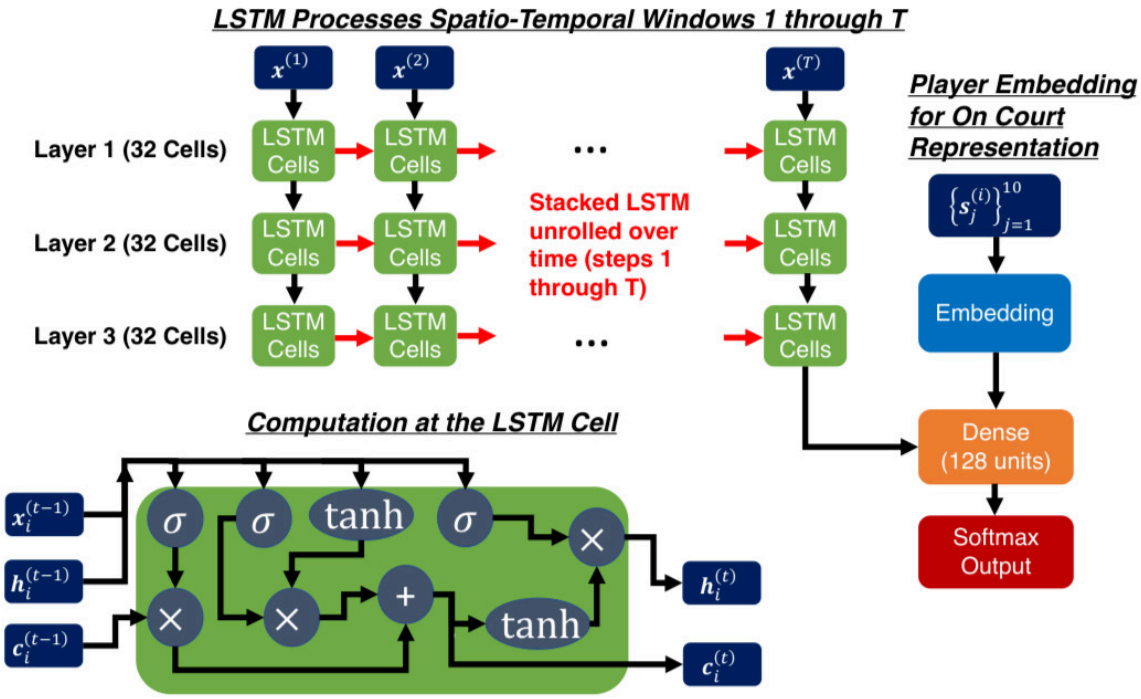

Figure 3: Architecture of the DeepHoops model.

3.1.2.1. Objective

Sicilia, Pelechrinis, and Goldsberry (2019) [[6]] address the challenge of evaluating the micro-actions that occur within basketball possessions by estimating, at each moment in time, the expected points remaining in a possession. Unlike traditional analytics, which emphasize discrete outcomes such as shots or turnovers, DeepHoops focuses on the evolving sequence of player and ball movements. The central idea is that each possession can lead to a variety of terminal actions—including field-goal attempts, shooting fouls, non-shooting fouls, turnovers, or null endings—and each of these terminal actions is associated with an expected point value. DeepHoops aims to predict the probability distribution over these terminal actions as a function of the spatio-temporal evolution of the play. By doing so, the model can quantify the value of micro-actions that traditionally go unmeasured but meaningfully influence possession outcomes. The overarching research objective is therefore to construct a deep learning architecture capable of capturing fine-grained spatio-temporal dynamics and translating them into a measure of possession value.

3.1.2.2. Methodological Framework

The foundational unit of analysis in DeepHoops is the possession, defined as a sequence of n moments, where each moment is encoded as a 24-dimensional feature vector [[6]]. The first 20 elements represent player coordinates; three additional elements store the (x, y, z) position of the ball; and the final element encodes the shot-clock value. With more than 134,000 tracked possessions, the authors extract for each moment a temporal window of size T leading up to time T.

To model temporal dynamics, the architecture employs a stacked LSTM module consisting of three layers with 32 cells each, relying on the Long Short-Term Memory mechanism introduced by Hochreiter and Schmidhuber [[7]]. This variant uses peephole connections, following Gers et al. [[8]], allowing cell states

to directly influence the gating functions. Optimization is performed using the Adam optimizer [[9]], and the network is trained using Backpropagation Through Time (BPTT) [[10]].

Because possession value depends on which players are involved, the architecture incorporates a separate embedding module that encodes lineup identity [[6]]. The LSTM representation and lineup embeddings are concatenated and passed to a softmax classifier predicting terminal event probabilities. These probabilities are used to compute the expected points remaining in the possession at time T.

3.1.2.3. Dataset

DeepHoops is trained on optical tracking data from 750 NBA games, captured at 25 frames per second using an advanced player- and ball-tracking system [[6]]. Each frame provides 3D coordinates for all players and the ball, producing millions of spatio-temporal samples. Event annotations—such as fouls, turnovers, and shots—allow the segmentation of each game into well-defined possessions and the labeling of terminal actions. This yields over 134,000 possessions with rich movement patterns and high-quality temporal alignment.

3.1.2.4. Models Implemented

The core model is the stacked LSTM architecture described above, leveraging principles from recurrent neural network design [[7]; [8]]. Three layers of 32 LSTM cells extract progressively richer temporal representations of possession dynamics. A lineup-embedding module introduces contextual information about personnel on the court. The concatenation of these components feeds into a fully connected network with a softmax output layer producing predicted terminal action probabilities.

3.1.2.5. Results and Conclusions

Model performance is evaluated using the Brier Score [[11]], a proper scoring rule for probabilistic predictions. Training proceeds for five epochs with a minimum improvement threshold of 0.01 [[6]]. Experiments vary the temporal window size K, with results showing that larger windows yield more accurate predictions. For K = 1, 2, 3, 4, the Brier Scores (BS), reference climatology scores (BSref), and Brier Skill Scores (BSS) all demonstrate performance improvements, with the best results at K = 4 (BS = 0.2659, BSS = 0.2114).

The authors also present reliability curves indicating that DeepHoops is well-calibrated. Beyond quantitative results, the study demonstrates that DeepHoops provides a powerful representation of evolving possession states, enabling analysts to assign value to micro-actions that traditional basketball statistics cannot measure.

	BS	BS_ref	BSS	Epoch Time (s)
K = 1	0.4569	0.6070	0.2472	2180
K = 2	0.3598	0.4920	0.2686s	2929
K = 3	0.3094	0.4017	0.2299s	3552
K = 4	0.2659	0.3371	0.2114	4200

Table 2: DeepHoops Brier Score (BS), Climatology Model Brier Score (BSref), and DeepHoops Brier Skill Score (BSS). DeepHoops outperforms the climatology (baseline) model in all cases. Performance is best for K = 2 (among the values examined). Epoch Time (in seconds) is lowest over all epochs

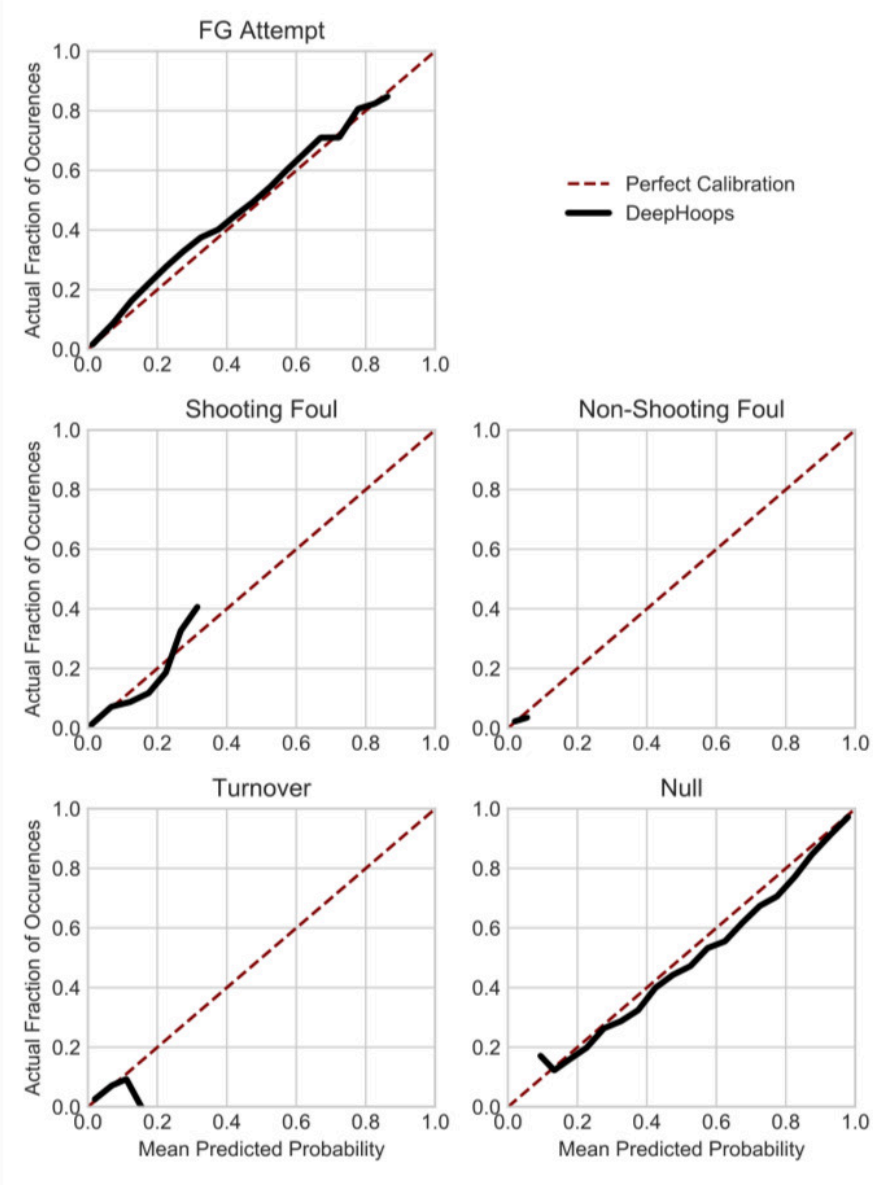

Figure 4: Reliability curves for DeepHoops' probability estimates. The dashed line y = x represents perfect calibration.

3.2. 3D Game Simulation

This category focuses on research efforts aimed at transforming traditional 2D sports imagery into enhanced 3D visual representations, enabling a richer and more immersive understanding of game dynamics. Soccer and basketball, in particular, pose significant challenges for such reconstruction due to rapid player motion, camera panning, occlusion, and the inherently limited depth cues of monocular broadcast footage. Researchers in this domain attempt either to simulate 3D structure from 2D images—for example, estimating depth maps or player geometry—or to reconstruct full 3D scenes that can be interacted with using augmented or virtual reality systems. These methods typically rely on advanced deep learning architectures capable of inferring spatial depth, body shape, motion, and camera pose directly from raw video input. In this section, we explore two representative studies: one focused on soccer, where

monocular video is converted into detailed 3D player models for AR visualization, and one centered on basketball, where spatio-temporal trajectories are used to generate predictive or reconstructed 3D representations. Together, these works highlight how deep learning has enabled a new generation of 3D modeling tools that enhance tactical analysis, fan engagement, and simulation capabilities in modern sports analytics.

3.2.1. Soccer On Your TableTop [[12]]

3.2.1.1. Objective

Rematas, Kemelmacher-Shlizerman, Curless, and Seitz (2018) [[12]] tackle the problem of transforming an ordinary monocular broadcast video of a soccer match into a full 3D reconstruction that can be viewed interactively using augmented reality (AR) devices. Traditional soccer broadcasts present the game from a fixed vantage point, significantly limiting the audience's spatial understanding of player motion, positioning, and tactical geometry. The authors aim to overcome these constraints by reconstructing both the players and the playing field in three dimensions, enabling users to examine the game from arbitrary viewpoints—including a tabletop AR environment. Their overarching research objective is therefore to design a system that uses deep learning, geometric reasoning, and camera pose estimation to recover per-player depth and geometry from a single 2D video, ultimately producing an immersive 3D visualization of real soccer footage.

3.2.1.2. Methodological Framework

A central methodological challenge is estimating the depth map of each player from a single 2D image, since monocular depth recovery is inherently ill-posed. To address this, the authors develop a dedicated depth-estimation neural network based on a sequence of eight hourglass modules, following the architecture introduced by Newell et al. [[13]]. The network takes as input a 256×256 cropped RGB image of a player and predicts a 64×64×50 volumetric representation encoding quantized depth.

The 50-channel output volume corresponds to 49 quantized depth bins plus one background class, with quantization defined relative to a vertical mid-plane passing through the player. Distances are discretized into 24 bins in front, 24 behind, and one at the plane itself, each representing 0.02 meters. The network is trained with an entropy loss for 300 epochs using a batch size of 6 [[12]].

Once depth is inferred, the authors integrate the depth estimation into a complete 3D reconstruction pipeline that includes camera pose estimation, player detection and tracking, and mesh generation. Estimated depth maps are projected into 3D and converted into detailed meshes, enabling full-scene 3D reconstruction viewable through AR.

3.2.1.3. Dataset

Because no dataset includes high-quality depth maps for real soccer players, the authors construct their own synthetic dataset using RenderDoc to intercept GPU calls from the FIFA video game engine [[12]]. By extracting Normalized Device Coordinates (NDC) and depth buffers, they obtain dense 3D point clouds for each player. After isolating players and discarding irrelevant geometry, they compile a dataset of approximately 12,000 RGB-depth pairs.

Although synthetic, the dataset is photorealistic and diverse thanks to FIFA's animation system, and provides the dense supervision necessary to train the depth-estimation network effectively.

3.2.1.4. Models Implemented

The core learning model is a CNN based on stacked hourglass modules [[13]], designed for dense prediction tasks requiring simultaneous global context and fine-grained spatial accuracy. The network repeatedly compresses and expands resolution, enabling it to infer depth from subtle visual signals like shading, limb configuration, and texture.

Beyond the network, the authors implement modules for camera pose estimation, player tracking, and 3D mesh construction, forming an end-to-end 3D reconstruction system capable of converting monocular video into AR-viewable content [[12]].

3.2.1.5. Results and Conclusions

Evaluation is conducted on a held-out test set of 32 RGB-depth pairs extracted from FIFA using the same RenderDoc method. The primary metric is the scale-invariant RMSE (st-RMSE). The authors compare their model to three baseline methods:

- A non-human generic depth estimation model [[14]]
- A human-specific depth estimation method [[15]]
- A parametric human-shape-model-fitting method based on SMPL [[16]]

Their method achieves an st-RMSE of **0.06**, outperforming the parametric SMPL model (0.14) and far surpassing both non-human (0.92) and non-soccer-trained (0.16) baselines. Foreground segmentation IoU reaches **0.86**, indicating accurate localization of player geometry.

The results show that the proposed depth-estimation network generalizes effectively from synthetic data to real broadcast video, allowing full 3D reconstruction of soccer matches for AR display. The study demonstrates that monocular soccer footage can be converted into compelling 3D visualizations, enabling new forms of analysis and fan engagement.

	st-RMSE	**IoU**
Non-human training	0.92	-
Non-soccer training	0.16	0.41
Parametric Shape	0.14	0.61
Their Model	0.06	0.86

Table 3: Results of Depth Estimation Network

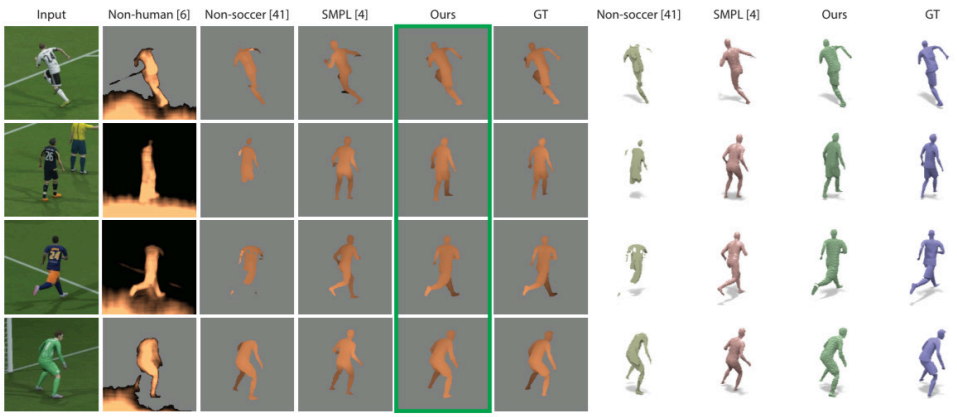

Figure 5: Results of the experiment

3.2.2. Basketball Trajectories [[17]]

3.2.2.1. Objective

Shah and Romijnders (2016) [[17]] investigate whether deep learning can be used to accurately predict the success of three-point shots in basketball, based solely on the spatio-temporal trajectory of the ball. Prior to this work, shot prediction models largely depended on contextual variables such as defender proximity, shooter identity, or game situation, rather than the actual physical arc of the ball. The authors frame the problem as a sequence-learning task, in which the flight path of the ball is treated as a time-series and the goal is to estimate, at each timestep, the probability that the shot will ultimately be made. Their main research objective is to determine whether a Recurrent Neural Network (RNN) equipped with Long Short-Term Memory (LSTM) units [[7]] can outperform traditional machine-learning models in predicting shot success by learning patterns directly from raw trajectory data. In doing so, the study aims to explore the viability of deep learning as a tool for physics-based sports analytics.

3.2.2.2. Methodological Framework

To capture the temporal dependencies inherent in ball flight, the authors employ a two-layer LSTM architecture using peephole connections, a variant introduced by Gers et al. [[8]] that allows each memory cell to inspect its internal state when computing gate activations. This design enables the model to learn fine-grained physical cues from the evolving trajectory. The input at each timestep consists of the ball's X, Y, and Z coordinates, along with the game clock value, providing both spatial and temporal context. As the time series progresses, the LSTM processes successive ball positions and outputs, at each timestep, the probability that the shot will be successful. These probabilities are computed by a softmax output layer and trained using cross-entropy loss.

The model is optimized using the Adam optimizer [[9]], which accelerates convergence through adaptive learning-rate adjustments. To examine the impact of engineered physics-based features, the authors also create a second input formulation that augments raw XYZ coordinates with additional variables derived from ball mechanics. These include per-timestep velocity components, distance to the rim, change in distance over time, and the shot angle relative to the rim. The expectation is that these physics-informed variables may help the model better infer the underlying ballistic dynamics of successful shots. The overall framework thus combines time-series modeling, physical intuition, and deep recurrent architectures to learn discriminative trajectory patterns.

3.2.2.3. Dataset

The study draws from the publicly available SportVU optical tracking system, which captures the positions of all players and the ball at 25 frames per second across all NBA arenas [[17]]. The dataset used for this research consists of over 20,000 three-point shot attempts collected from 631 games during the early portion of the 2015–2016 NBA season. The system records precise spatial coordinates of the ball in three dimensions—X along the court length, Y along the width, and Z representing height—making it ideally suited for trajectory-based modeling. Of the collected shots, 35.7% were made, providing a realistic class distribution for predictive modeling.

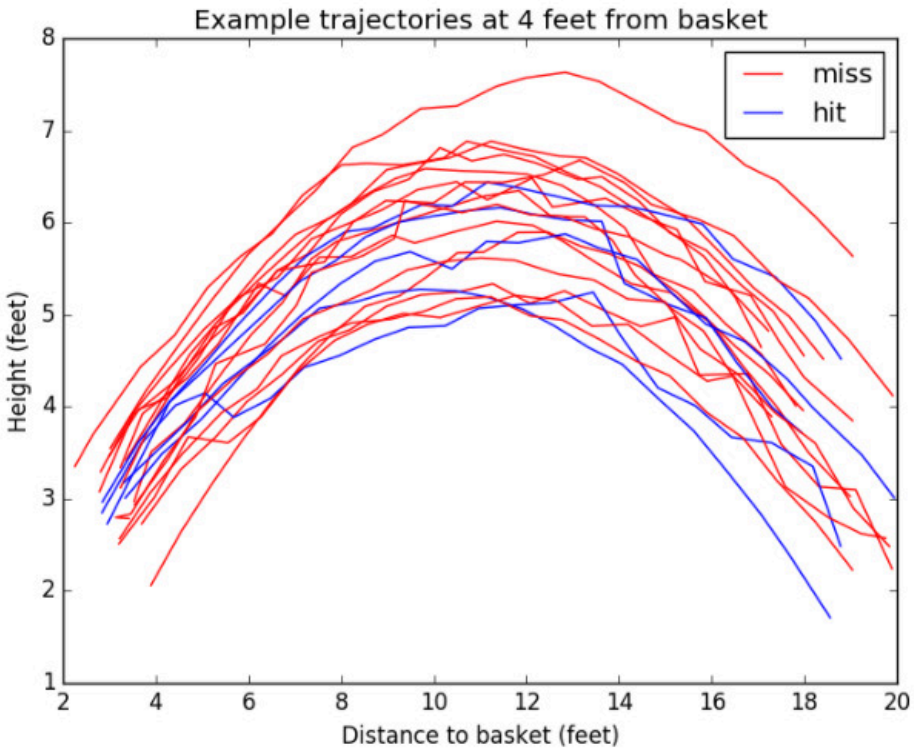

Figure 6: basketball data examples

To construct the training data, each three-point attempt is represented as a sequence of XYZ coordinates indexed over time, along with the game clock. A secondary dataset is derived by augmenting these trajectories with engineered features inspired by projectile dynamics. These include the ball's per-axis movement differences between frames, distance from the ball to the center of the rim, temporal differences in that distance, and the angle of approach. By combining raw positional data with physics-based attributes, the authors create two distinct model inputs to evaluate whether deep networks can learn these dynamics implicitly or whether engineered features provide measurable predictive improvements.

3.2.2.4. Models Implemented

The primary model used in the study is a Recurrent Neural Network with Long Short-Term Memory (LSTM) units [[7]], designed to capture the temporal evolution of the ball's flight path. The network consists of two stacked LSTM layers incorporating peephole connections [[8]], enabling it to learn temporal patterns and internal cell-state dynamics more effectively than standard LSTMs. The final output is a probability estimate of shot success produced at each timestep. The learning process utilizes cross-

entropy loss and Adam optimization [[9]], making the model well suited for binary classification tasks involving sequential data.

For comparison, the authors also implement two baseline machine-learning models:
- a Generalized Linear Model (GLM) [[18]], and
- Gradient Boosted Machines (GBM) [[19]],

to evaluate the advantages of deep sequence modeling over traditional approaches.

3.2.2.5. Results and Conclusions

The data is split into an 80–20 train-test partition, and model performance is evaluated using the Area Under the ROC Curve (AUC), a standard metric for binary classification introduced by Hanley and McNeil [[20]]. The baseline GLM performs poorly, achieving an AUC of 0.53, only slightly above random guessing. The GBM model performs significantly better with an AUC of 0.80, indicating that engineered features and non-linear decision boundaries are effective. However, the LSTM-based RNN achieves the highest performance with an AUC of 0.843, demonstrating that deep recurrent architectures are particularly well suited for capturing the temporal and physical nuances of ball trajectories.

The authors conclude that RNNs outperform traditional models in trajectory-based prediction tasks by learning the dynamics of ball motion directly from raw spatio-temporal sequences. Their results suggest that deep learning can detect subtle cues such as arc height, release angle, velocity decay, and rim-approach angle—factors strongly associated with shot success. More broadly, the study highlights the potential for deep sequence models to advance sports analytics by leveraging high-resolution trajectory data, offering a framework that could be extended beyond basketball to other domains where physics-driven motion plays a central role.

	GLM	**GBM**	**RNN**
AUC	0.53	0.80	0.843

Table 4: Results of Experiment

3.3. Game Action Categorization

This category focuses on the problem of automatically recognizing and classifying actions within soccer video sequences, an area that has attracted substantial research attention due to the fast-paced and visually complex nature of the sport. Action recognition aims to identify meaningful soccer-specific events—such as free kicks, fouls, goals, offsides, corner kicks, or player interactions—directly from video footage, without requiring manual annotation or external metadata. These actions often involve subtle temporal dynamics, rapid player movement, and significant camera motion, making them challenging to detect using traditional vision methods. Modern approaches therefore rely on advanced machine learning techniques capable of modeling both spatial and temporal information present in video streams. In this section, we examine two representative research works that address this challenge using deep learning, each proposing a different methodology for extracting frame-level features, modeling temporal dependencies, and generating accurate action classifications. Together, these studies demonstrate how recurrent architectures, motion descriptors, and visual-content representations can be combined to effectively interpret and categorize soccer actions from raw video sequences.

3.3.1. Action Classification using LSTM RNN [[21]]

3.3.1.1. Objective

Baccouche et al. (2010) [[21]] investigate the challenge of classifying human actions directly from video sequences in the context of soccer, using only visual content without relying on handcrafted priors or domain-specific rules. Prior to this work, many action-recognition approaches depended on additional cues such as temporal segmentation rules or motion templates, which limited their adaptability. The authors instead propose a method that learns action representations directly from raw video content, treating the video as a sequential signal composed of evolving frames. Their research objective is to develop a system capable of recognizing complex soccer actions using only extracted frame descriptors and a recurrent architecture capable of modeling temporal evolution. By doing so, they aim to demonstrate that recurrent neural networks, particularly LSTM-based models [[7]], can effectively capture temporal patterns in sports videos and outperform classical baselines that rely on static image descriptors.

3.3.1.2. Methodological Framework

The authors follow a sequential learning framework in which each video is divided into frames, and each frame is transformed into a visual descriptor capturing appearance and motion. These descriptors are then fed into a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [[7]] that processes them timestep by timestep, learning how frame content evolves as the action unfolds. The final classification decision is produced by aggregating frame-level predictions.

fig. 7 show's the approach.

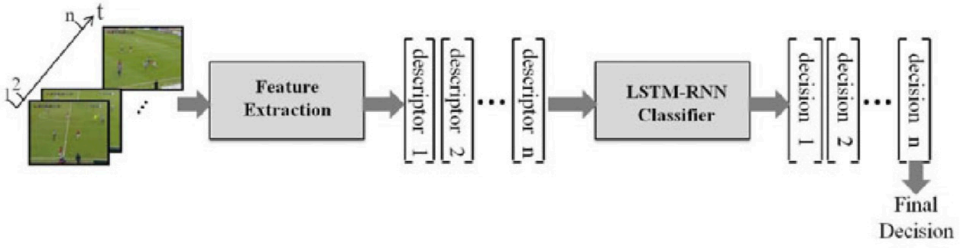

Figure 7: approach used by researchers

To represent visual content, the authors extract a Bag-of-Words (BoW) descriptor for every frame, following approaches such as Video Google [[22]]. To complement appearance information, they compute dominant motion features using a SIFT-based approach [[23]]. SIFT keypoints are extracted from consecutive frames and matched using a KD-tree structure. Since raw matches contain both camera and player motion, RANSAC [[24]] is applied to eliminate outliers and isolate global camera motion, while static elements such as TV logos are removed. These combined appearance and motion descriptors serve as input to the LSTM model.

3.3.1.3. Dataset

All experiments are conducted using the MICC-Soccer-Actions-4 dataset, a benchmark designed for soccer action classification. It contains annotated video clips corresponding to four soccer-specific action categories, recorded from real broadcast footage and manually segmented into distinct events. Each clip is labeled with the action type and contains consistent camera angles typical of televised soccer matches. Although relatively small by modern standards, the dataset offers a challenging test bed due to variations in lighting, occlusions, player density, and rapid camera motion.

3.3.1.4. Models Implemented

The primary model is an LSTM-RNN [[7]] designed to capture long-term dependencies across frames. Classical RNNs suffer from vanishing gradients, making them unsuitable for long videos. LSTMs overcome this via their Constant Error Carousel (CEC) and gating mechanisms. The network consists of one hidden recurrent layer with a dimensionality determined by the input descriptor, and a SoftMax output layer predicting action probabilities at each timestep. The authors use 150 LSTM cells, noting that larger models risk overfitting while smaller ones may fail to capture relevant temporal patterns.

Training uses Online Backpropagation Through Time (BPTT) [[10]] with a learning rate of $10^{\{-4\}}$ and momentum 0.9. The combination of BoW appearance features, SIFT–RANSAC motion descriptors, and recurrent modeling enables the system to learn richer temporal dynamics than conventional classifiers such as k-NN [[25]] or SVM [[26]].

3.3.1.5. Results and Conclusions

Experiments use a 3-fold cross-validation scheme on the MICC-Soccer-Actions-4 dataset. Using BoW features alone, a k-NN classifier achieves 52.75% [[25]], while an SVM improves to 73.25% [[26]], indicating that appearance provides meaningful but incomplete discriminative power. Incorporating temporal modeling via an LSTM-RNN increases accuracy to 76%.

When dominant motion features serve as input to the LSTM, performance rises to 77%. The most significant gain occurs when both BoW and dominant motion features are combined, achieving 92% accuracy —far surpassing all baselines. Confusion matrices show that the combined model reduces misclassification between visually similar actions, confirming the importance of modeling both appearance and temporal dynamics.

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0.92**	0.08	0	0
Placed-kick	0.08	**0.8**	0	0.12
Shot-on-goal	0	0.2	**0.72**	0.08
Throw-in	0.12	0.12	0.16	**0.6**

(a)

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**0.64**	0.28	0.08	0
Placed-kick	0.08	**0.68**	0.08	0.16
Shot-on-goal	0.08	0	**0.88**	0.04
Throw-in	0.08	0	0.04	**0.88**

(b)

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**1**	0	0	0
Placed-kick	0.04	**0.84**	0.08	0.04
Shot-on-goal	0	0.12	**0.88**	0
Throw-in	0.04	0	0	**0.96**

(c)

Figure 8: Confusion matrices : (a) - BoW-based approach (b) - Dominant motion-based approach (c) - Combination

below is the table that showcase the results

	Classification Rate
BoW + k-NN	52.75%
BoW + SVM	73.25%
BoW + LSTM-RNN	76%
Dominant motion + LSTM-RNN	77%
BoW + dominant motion + LSTM-RNN	92%

Table 5: Results of Experiment

3.3.2. Soccer Video Summarization using Deep Learning [[27]]

3.3.2.1. Objective

Agyeman, Muhammad, and Choi (2019) [[27]] address the challenge of automatically identifying and extracting important moments from full-length soccer matches. Because broadcast soccer videos are long and contain large stretches of low-activity gameplay, manually locating highlights such as goals, shots, or key build-ups is time-consuming for analysts, coaches, and viewers. The authors frame the problem as one of soccer-specific video summarization: given a full match video, the goal is to automatically detect salient soccer actions and assemble them into a concise highlight summary. Their research objective is to build a deep-learning-based pipeline that learns relevant spatio-temporal patterns from raw video data, identifies clips that constitute meaningful events, and produces a high-quality, short-form summary without reliance on handcrafted rules or manual intervention.

3.3.2.2. Methodological Framework

To achieve this, the authors introduce a two-stage deep neural pipeline combining a 3D Convolutional Neural Network (3D-CNN) with a recurrent sequence model. The first stage uses a 3D-CNN based on a Residual Network (ResNet) backbone [[4]], leveraging 3D convolutions to jointly capture spatial structure and temporal motion cues within short video segments. Because 3D convolutions naturally encode appearance, movement, and short-term dynamics, they provide a learned alternative to handcrafted motion descriptors. The 3D-CNN is trained on manually annotated clips to classify soccer-specific action categories, effectively functioning as an action-recognition module.

Once trained, the 3D-CNN produces high-level spatio-temporal feature embeddings for segments extracted from full match videos. These embeddings are passed to an LSTM-based recurrent network [[7]], which models long-range temporal dependencies and contextual information across entire matches. The LSTM processes sequences of clip-level features and outputs a “highlight relevance score” for each segment. This relevance score determines whether the segment contributes important information to the final summary. ReLU activations [[5]] and fully convolutional principles [[2]] further support the representation learning throughout the architecture. Segments with the highest predicted relevance are selected to construct the final summarized video, allowing the system to generate a coherent, concise highlight reel.

3.3.2.3. Dataset

To train the 3D-CNN action recognizer, the authors construct a dataset of 744 manually annotated soccer video clips extracted from broadcast match footage. Each clip is labeled into one of five soccer-action classes corresponding to key events of interest for highlight generation. These labeled clips form the supervised training set for learning spatio-temporal representations specific to soccer gameplay.

For evaluation, the authors test their summarization system on ten full-length soccer match recordings not used during training. These complete matches allow the pipeline to be assessed on realistic, untrimmed, and variable-length videos. This two-tier dataset design — short action-level clips for training and full matches for evaluation — enables the model to learn meaningful soccer events while also demonstrating practical summarization performance on real broadcast videos.

3.3.2.4. Models Implemented

The implemented system consists of two main deep networks. The first is a 3D-CNN with a ResNet-style backbone [[4]], which learns rich spatio-temporal representations from short clips. Unlike 2D CNNs, the 3D model captures both motion and appearance simultaneously, making it well suited for action recognition. The second component is an LSTM-RNN [[7]] that receives sequential clip-level features and learns temporal context across an entire match. This recurrent stage is critical because soccer highlights often depend on buildup, transitions, and evolving play structure. The combination of short-term 3D convolutional modeling and long-term recurrent modeling produces a robust pipeline for identifying important video segments. After classification and scoring, a selection mechanism extracts segments with the highest predicted highlight relevance and stitches them into a summary video.

3.3.2.5. Results and Conclusions

To evaluate performance, the authors generated highlight summaries for ten full soccer matches and conducted a user study with 48 participants from eight countries. Participants rated the summaries on a Mean Opinion Score (MOS) scale, and the system achieved an average score of 4 out of 5 points. This indicates that viewers found the generated summaries to be of high quality, engaging, and representative of important match events. The results support the conclusion that deep-learning-based summarization — driven by 3D CNN action modeling and LSTM temporal reasoning — can successfully automate the highlight-creation process for soccer videos.

The authors further argue that their architecture is generalizable to other sports or long-form video domains, provided domain-specific annotated clips are available for training. This positions the method as a promising step toward automated, content-aware sports video summarization.

4. Limitations

While the studies reviewed in this survey highlight meaningful advances in sports analytics through deep learning, they collectively exhibit several important limitations that constrain their generalizability, robustness, and practical deployment. A common limitation across all categories is the dependency on highly controlled or domain-specific datasets, which restricts model applicability to broader real-world scenarios. For instance, many action-recognition and prediction models rely on curated datasets with clean annotations, stable broadcast viewpoints, or synthetic training data extracted from game engines. These controlled conditions differ significantly from the variability found in real competitive environments, where occlusions, camera shifts, lighting changes, and inconsistent broadcast angles introduce noise that many models may not be equipped to handle.

In the Game Prediction category, models such as SoccerMap and DeepHoops rely heavily on dense spatio-temporal tracking data, which is expensive, proprietary, and unavailable for most leagues outside top-tier competitions. Their predictions are therefore limited to environments where high-quality optical tracking is deployed. Moreover, these models often treat micro-actions in isolation and may not fully capture tactical dependencies, inter-team dynamics, or long-term strategic behaviors. Their interpretability also remains limited: although SoccerMap provides visually intuitive probability maps, underlying model decisions are still influenced by complex feature interactions that remain opaque.

In the Game Action Categorization category, models that classify actions from video sequences frequently suffer from small dataset size, limited action diversity, and narrow domain focus. Datasets like MICC-Soccer-Actions-4 include only a handful of action classes and lack representation of real-world variability. Many approaches also rely on handcrafted features—such as SIFT-based dominant motion or Bag-of-Words descriptors—that may not generalize well across different camera viewpoints or broadcast styles. Even recurrent architectures, while effective at capturing temporal dynamics, may struggle with long-range temporal contexts or be susceptible to overfitting when trained on limited video corpora.

Within the 3D Game Simulation category, limitations are often tied to monocular depth ambiguity and reliance on synthetic datasets. Methods like "Soccer on Your Tabletop" depend on depth maps extracted from a video game engine rather than real match footage, creating a domain gap that can undermine performance on real-world videos. Reconstructing high-quality 3D meshes from 2D broadcast images also remains extremely challenging due to occlusions, low resolution, heavy compression, and unpredictable camera motion. Similarly, reconstruction models may fail in crowded scenes or under conditions where player silhouettes overlap, preventing accurate depth estimation or mesh generation.

Across all categories, a major technical limitation is the lack of multimodal integration. Most models depend on a single type of input—either tracking data, raw video, or isolated trajectory coordinates—rather than combining visual cues, contextual game information, player identity, or tactical metadata. This restricts their ability to understand complex game situations. Evaluation metrics across studies are also inconsistent, ranging from AUC and Brier Score to st-RMSE and classification accuracy, making direct cross-study comparison difficult.

Finally, there is a notable absence of real-time performance analysis and deployment feasibility across nearly all works. Many architectures are computationally heavy, with large LSTM stacks, hourglass networks, or 3D CNN pipelines that may be impractical for live analytics during matches. Consequently, while the surveyed studies demonstrate promising research directions, significant work remains to improve scalability, generalization, multimodal integration, and real-world robustness before these methods can be universally adopted in professional sports environments.

Bibliography

[1] J. Fernández and L. Bornn, "SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer," *arXiv preprint arXiv:2010.10202*, 2020.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[3] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained Convolutional Neural Networks for Weakly Supervised Segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1796–1804.

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

[5] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[6] A. Sicilia, K. Pelechrinis, and K. Goldsberry, "DeepHoops: Evaluating Micro-Actions in Basketball Using Deep Feature Representations of Spatio-Temporal Data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2096–2104.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, 2000.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2015.

[10] R. J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," *Neural Computation*, vol. 1, pp. 270–280, 1989.

[11] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[12] K. Rematas, I. Kemelmacher-Shlizerman, B. Curless, and S. Seitz, "Soccer on Your Tabletop," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4738–4747.

[13] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network," in *Advances in Neural Information Processing Systems*, 2014.

[15] P. Ghosh, D. Tzionas, and M. J. Black, "Learning Human Motion Models for Long-Term Predictions," in *CVPR Workshops*, 2017.

[16] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic Estimation of 3D Human Shape and Pose from a Single Image," in *European Conference on Computer Vision*, 2016, pp. 561–578.

[17] R. Shah and R. Romijnders, "Applying Deep Learning to Basketball Trajectories," *arXiv preprint arXiv:1608.03793*, 2016.

[18] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Chapman, Hall, 1989.

[19] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2001.

[20] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[21] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks," in *International Conference on Artificial Neural Networks*, Springer, 2010, pp. 154–159.

[22] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[23] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[24] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[25] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[26] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[27] R. Agyeman, R. Muhammad, and G. S. Choi, "Soccer Video Summarization Using Deep Learning," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 270–273.