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 constitutes the principal body of the survey. The research works examined are organized into three major thematic categories: Game Prediction, Game Action Categorization, and 3D Game Simulation. Each category is subdivided into multiple subsections, with each subsection dedicated to a single study included in the survey.

For every research work, we provide a structured and critical summary that encompasses:

- the problem formulation and research objectives,
- the methodological framework adopted by the authors,
- the datasets utilized, including details on their acquisition and construction,
- the models or algorithms implemented, and
- the empirical results and conclusions reported.

This organizational structure ensures clarity, coherence, and comparability across the diverse set of studies reviewed.

3.1. Game Prediction

In this category researches focused on predicting probabilities of micro actions in game that might occur using different parameters such as player positions, ball positions, linup etc. This section is composed of two researches one for soccer and another for basketball and will summarize their approach the problem they solved, the dataset they used, the models they used and of course the results that were obtained along with the metrics that were used to measure them.

3.1.1. SoccerMap [LINK]

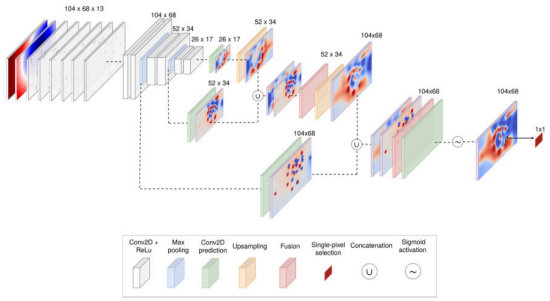

Figure 1: Architecture of the SoccerMap model.

3.1.1.1. Objective

SoccerMap aims to estimate continuous probability surfaces representing potential passing locations during a soccer match. The method enables coaches and analysts to visually inspect and understand player positioning, decision-making tendencies, and tactical structures throughout a game.

3.1.1.2. Methodological Framework

The model uses convolution layers to pick up meaningful patterns from the soccer field map at different zoom levels while keeping the image size unchanged and avoiding border issues. It downsamples the map with pooling to understand broader, more general patterns, then upsamples it smoothly to regain detail without creating visual artifacts. A fully convolutional design allows the model to make a prediction at every location on the field rather than just one final output. Finally, small 1×1 filters are used to turn these learned features into passing-probability predictions at each spot, and information from different zoom levels is combined so the model benefits from both fine and coarse details.

3.1.1.3. Dataset

high frequency, spatio temporal data is used. tracking data extracted from videos of soccer match consisting of 2d locations of players and the ball at 10 frames per second, allong with manually tagged passes, they arranged the data in a matrix of c x l x h, where l and h are the high level field coordinates and c represents the low level parameters that are passed to this architecture. low level features are calculated with things like liklehood of nearby teammates

they used tracking data, and event data from 740 English Premier League matches from the 2013/2014 and 2014/2015 season, provided by STATS LLC [LINK].

3.1.1.4. Models Implemented

they used a combination of models combining them in a single architecture, they are mentioned in fig.1

3.1.1.5. Results and Conclusions

They split the data into 60:20:20 split which means 60% training data, 20% validation data, 20% test data.

They benchmarked against two models. Logistic Net and Dense2 Net. Logistic Net consists of a single *sigmoid* [KEYWORD] activation unit while Dense2 is *Neural network* [KEYWORD] with two dense layers followed by *ReLu* [KEYWORD] [LINK] activation unit and a sigmoid output unit.

The Metrics they used are *Negative Log Loss* [KEYWORD] and *ECE* [KEYWORD] for every model

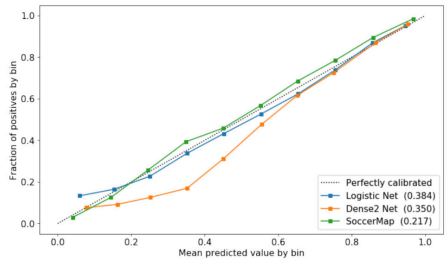

Figure 2: Visual comparison of model outputs on pass-probability surfaces.

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 [LINK]

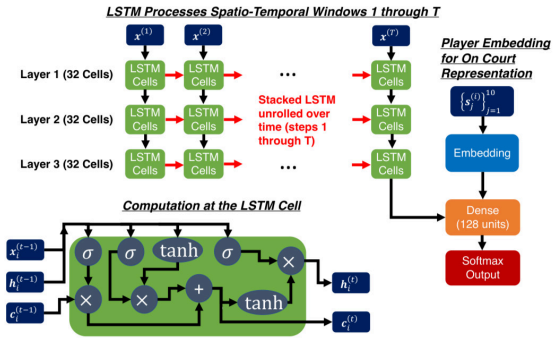

Figure 3: Architecture of the DeepHoops model.

3.1.2.1. Objective

They created a Neural Network Architecture called DeepHoops which analyses spatio temoporal data of NBA games and predicts expected points to be scored as progression progresses. They estimate the probabiliites of terminal actions (e.g take goal, foul, turnover), each of these terminal actions is associated with an expected point value.

3.1.2.2. Methodological Framework

they used a concept of possessions which is a sequence of n moments where each moment is a 24-dimenstional vector. the first 20 capture the location of 10 players (x, y). the next three represent the court location and height of the ball (x, y, z), the last elements represents the current value of shot clock.

their dataset consisted of more than 134,000 such possessions of interest.

they define a temporal window which acts as input to the DeepHoops architecture. window defined at a moment tao captures T moments up to a time of interest.

each window is labelled with an outcome that represents the type of terminal action that occured at the end of the window.

set of terminal actions included

- field goal attempts
- shooting foul
- Non-shooting foul
- turnover
- null

the architecture is mentioned above in fig 3. they used two modules one was a stacked *LSTM* [LINK] [KEYWORD] network which was responsible in learning representation of the data up to time tao, this allowed for important information about on-court actions. the additional module was used to model who is on the courl to assess the impact of different lineups

they use 32 LSTM cells for each 3 layers.

3.1.2.3. Dataset

they used optical tracking data of 750 NBA games, which represents NBA games as athree dimensional coordinate system. the data they used was highly annotated and allowed labelling.

3.1.2.4. Models Implemented

They used a stacked LSTM network. mentioned in fig 3

3.1.2.5. Results and Conclusions

The metric they used to calculate the accuracy was *Brier Score* [KEYWORD] [LINK] over 5 *epochs* [KEYWORD] with minimum *improvement rate* [KEYWORD] of 0.01

	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 1: 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

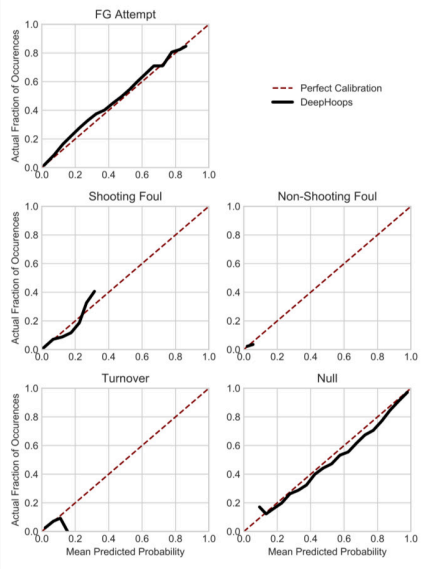

Figure 4: reliability Curves for DeepHoops' probability estimates. The dashed line y = x represents perfect calibration

3.2. 3D Game Simulation

In this category researchers tried to simulate 2D images or tried to completely reconstruct the 2D images in 3D models. This section will contain 2 researches one for soccer one for basketball following the previous pattern.

3.2.1. Soccer On Your TableTop [LINK]

3.2.1.1. Objective

The objective of the researchers was to transform a monocular video of a soccer game into a 3d reconstruction, in which players would be rendered interactively with an Augmented Reality Device.

3.2.1.2. Methodological Framework

A key component for the researchers was to estimate the *depth map* [KEYWORD] ofa particular player.

They used a depth estimation neural network. the input was a 256x256 RGB cropped image. The input is processed by a series of 8 *hourglass modules* [KEYWORD] [LINK] and the output is a 64x64x50 volume. representing 49 quantized depths and 1 background class.

The model was trained with entropy loss with a batch size of 6 for 300 epochs.

they estimate a virtual vertical plane passing through the middle of the player and calculate its depth w.r.t. the cam- era. Then, we find the distance in depth values between a player's point and the plane. The distance is quantized into 49 bins (1 bin at the plane, 24 bins in front, 24 bins behind) at a spacing of 0.02 meters, roughly covering 0.5 meters in front and in back of the plane (1 meter depth span).

then they follow a pipeline to reconstruct the full 3d game

which includes camera Pose Estimation then Player Detection and Tracking and then mesh generation this way they are able to get a 3d reconstruction of the orignal 2d game.

3.2.1.3. Dataset

researchers aquire the dataset of depthmaps for players by intercepting GPU calls between the game engine and the GPU of the game FIFA using RenderDoc. they aquired the *NDC* [KEYWORD] (Normalized Device Coordinates). This process gives them a point cloud and after removing everything but the player they collected 12000 image-depth pairs

3.2.1.4. Models Implemented

they used a *CNN* [KEYWORD] [LINk] with hour glass modules to estimate the depth buffers for 2d images.

3.2.1.5. Results and Conclusions

They evaluate their performance using a held-out datasets from FIFA video game captures. the data was created using the same process as in training data and contained 32 RGB depth pairs of images. The metric they used to measure their performance was (scale invariant- Root Mean Squared) *st-RMSE* [KEYWORD].

They compare with three different approaches

1. non human-specific depth estimation [LINK]
2. human-specific depth estimation [LINK]
3. fitting a parametric human shape model to 2D pose estimation [LINK]

	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 2: Results of Depth Estimation Network

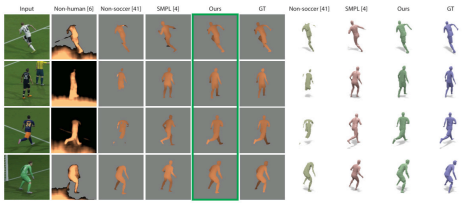

Figure 5: results of the experiment

3.2.2. Basketball Trajectories [LINK]

3.2.2.1. Objective

This research uses Recurrent Neural Networks *(RNN)* [KEYWORD] [LINK] to predict whether a 3 point shot would be successfull or not

3.2.2.2. Methodological Framework

Popular variant of RNN with long-short term memory (LSTM) is used. the network architecture relies on a two layered LSTM using peephole connections. the input to the LSTM is the XYZ data and the game clock. at each time step RNN predicts the probabilitiy of a successfull shot. the probability comes from a softmax layerand is trained based on cross entropy error.

An Adam optimizer is used in the model.

3.2.2.3. Dataset

the dataset used in the study stems from the publically availaible SportsVu dataset. SportVu is an optical tracking system installed by the National Basketball Association (NBA) in all 30 courts to collect real-time data. The tracking system records the spatial position of the ball and players on the court 25 times a second during a game.

The dataset consisted of over 20,000 three point shots attempts from 631 games. the data was taken from the NBA site in the beginning of 2015-2016. The percentage of made shots in the data set is 35.7%. Fig 5 shows the example of the dataset used.

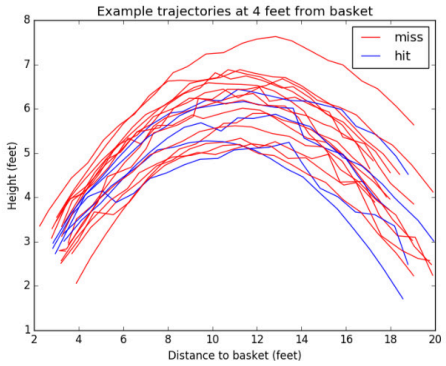

Figure 6: basketball data examples

The first dataset consists of only the X, Y, Z, and game clock variables representing the location of the ball in three dimensions over time. X refers to the length of the court, Y is the width of the court, and Z is the height of the ball. A second dataset is created with additional variables based on the physics of ball trajectories. The belief was that these variables would add more information over just the location data for machine learning models. Specifically, the added variables included the difference in movement over each time period for each dimension. Three other variables included: the distance to the center point of the rim, the difference over time for this distance, and the angle of the ball with respect to the rim.

3.2.2.4. Models Implemented

The research uses RNN (Recurrent Neural Networks) with the long short-term memory(LSTM) units

3.2.2.5. Results and Conclusions

The data is split in a 80:20 split for training and test respectively.

The metric they used to measure their accuracy was *AUC* [KEYWORD] [LINK]. they build classifiers using a Generalized Linear Model (GLM) and gradient boosted machines (GBM)

The below table showcase the results of the model with the baseline models.

	GLM	GBM	RNN
AUC	0.53	0.80	0.843

Table 3: Results of Experiment

3.3. Game Action Categorization

This cateogry focuses on classifying video sequences based on the actions that are being performed inside them. Examples for such actions in soccer would be free kick, red card/yellow card, goal, outside etc. In this section as always we will mention 2 researches focusing on this area providing a summary

3.3.1. Action Classification using LSTM RNN

3.3.1.1. Objective

The main problem the researches are trying to solve is how can a video sequence which is just a series of frames be classified according to the actions that are beign peformed in the video. They picked Soccer to conduct this research on. They relied solely on the visual content analysis to classify different actions which is different from previous approaches who utilized prior knowledge to classify actions.

3.3.1.2. Methodological Framework

For every video, we divide it into frames, and each frame is converted into a descriptor (one descriptor per image). Then we train an LSTM-RNN to predict which action is being performed. These descriptors change over time according to the frames. The final decision is made by combining all frame-level decisions.

fig. 7 show's the approach.

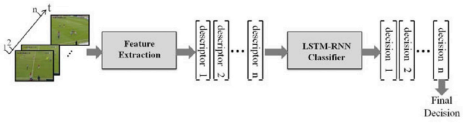

Figure 7: approach used by researchers

Features are extracted in the following way :-

1. Visual content representation: A Bag of Words approach BoW is a method that recognizes objects using a histogram of visual words (meaning a pattern that repeats across many images). One BoW is taken for each frame.
2. A SIFT-based approach for Dominant Motion Estimation Researchers added an extra feature called dominant motion, which captures movement in the video, especially the camera's movement. They extract SIFT feature points from two consecutive frames, then match these points (using a KD-tree) to understand the motion. TV logos, which have no motion, are removed. RANSAC is used so that only camera motion remains while ignoring players' random movement.

3.3.1.3. Dataset

they used the MICC-Soccer-Actions-4 Dataset.

3.3.1.4. Models Implemented

Action classification using *LSTM-RNN*[KEYWORD][LINK] is done by feeding each frame's descriptor to the network timestep-by-timestep, where the LSTM, an improved version of RNN, handles long term information using CEC (Constant Error Carousel) and gates that decide what to store or discard, overcoming the issue of RNNs forgetting old information in long sequences. The network architecture consists of one hidden RNN layer whose size depends on the input features, and a SoftMax layer at the output to make predictions at each timestep. A total of 150 LSTM cells are used more can cause *overfitting* [KEYWORD], while fewer may prevent proper learning and the model is trained using Online *BPTT* [KEYWORD] [LINK] with a *learning rate*[KEYWORD] of 10^{-4} and momentum 0.9.

3.3.1.5. Results and Conclusions

all the experiments conducted by the researchers were carried out on MICC- Soccer-Actions-4 dataset [LINK] with a *3-fold cross validation scheme* [KEYWORD]

Fig 8 showcases the confusion matrices of different approaches

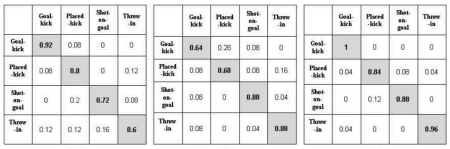

Figure 8: Confusion matrices : (a) - BoW-based approach (b) - Dominant motion-based approach (c) - Combination of the BoW and the dominant motion

below is the table that showcase the results

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

Table 4: Results of Experiment