Classifying Soccer Ball-on-Goal Position Through Kicker Shooting Action

Javier Torón Artiles[a], Daniel Hernández-Sosa[b], Oliverio J. Santana[c], Javier Lorenzo-Navarro[d] and David Freire-Obregón[e]
SIANI, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Keywords: Computer Vision, Soccer, Free Kick, Human Action Recognition, Dataset.

Abstract: This research addresses whether the ball's direction after a soccer free-kick can be accurately predicted solely by observing the shooter's kicking technique. To investigate this, we meticulously curated a dataset of soccer players executing free kicks and conducted manual temporal segmentation to identify the moment of the kick precisely. Our approach involves utilizing neural networks to develop a model that integrates Human Action Recognition (HAR) embeddings with contextual information, predicting the ball-on-goal position (BoGP) based on two temporal states: the kicker's run-up and the instant of the kick. The study encompasses a performance evaluation for eleven distinct HAR backbones, shedding light on their effectiveness in BoGP estimation during free-kick situations. An extra tabular metadata input is introduced, leading to an interesting model enhancement without introducing bias. The promising results reveal 69.1% accuracy when considering two primary BoGP classes: *right* and *left*. This underscores the model's proficiency in predicting the ball's destination towards the goal with high accuracy, offering promising implications for understanding free-kick dynamics in soccer.

1 INTRODUCTION

In the 2021/22 season, the top 20 revenue-generating clubs collectively made a profit of €9.2 billion, marking a 13% increase from the previous season and nearly reaching the pre-pandemic levels of 2018/19. This resurgence was driven by the return of fans to stadiums, resulting in a significant increase in matchday revenue, which rose from €111 million to €1.4 billion. The revenue composition of clubs in 2021/22 returned to pre-pandemic levels, with 15% from matchday activities, 44% from broadcasting, and 41% from commercial sources (Deloitte, 2023). Furthermore, the data indicates that the 2022 FIFA World Cup, held in Qatar, garnered the highest viewership in the tournament's history, with over five billion spectators tuning in through diverse platforms, surpassing more than half of the global population (FIFA, 2022).

This remarkable financial, as well as the widespread global viewership of soccer events, underscore the tremendous potential and impact of soccer as a mass sport. Furthermore, the evolution of soccer continues after these outstanding statistics. The introduction of technology into the sport is emerging as a pivotal factor, shaping both its on-field dynamics and off-field engagement. According to Microsoft, during a match, players navigate the entire field at high speed, necessitating the deployment of up to 16 fixed cameras for optical tracking positioned around the perimeter of each stadium, capturing a staggering 3.5 million data points per game (Microsoft, 2023). This data is subsequently processed through the Mediacoach platform, making it accessible to clubs and fans through match broadcasts and digital content. Microsoft also remarks that the data strategy is designed to give clubs invaluable insights for adapting training schedules, scrutinizing opponents, and preparing for match days.

In this context, the integration of technology into soccer has brought about a significant transformation in how the sport is played, assessed, and enjoyed. Several studies and technological innovations have highlighted the potential of technology to enhance

[a] https://orcid.org/0009-0000-5082-310X
[b] https://orcid.org/0000-0003-3022-7698
[c] https://orcid.org/0000-0001-7511-5783
[d] https://orcid.org/0000-0002-2834-2067
[e] https://orcid.org/0000-0003-2378-4277

Artiles, J., Hernández-Sosa, D., Santana, O., Lorenzo-Navarro, J. and Freire-Obregón, D.
Classifying Soccer Ball-on-Goal Position Through Kicker Shooting Action.
DOI: 10.5220/0012417100003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 79-90
ISBN: 978-989-758-684-2; ISSN: 2184-4313

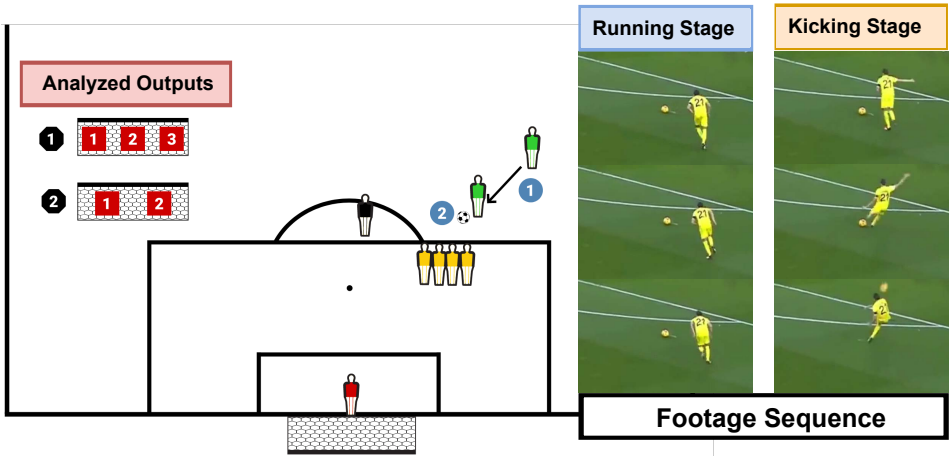

Figure 1: **Free-Kick BoGP Classification.** Our proposal involves a thorough analysis of free-kick actions by integrating data from various sources, including free-kick metadata and HAR embeddings. Critically, our classifier combines contextual information with the two-stream action recognition embeddings to make accurate predictions regarding the ball's placement concerning the goal. It is important to note that these experiments relied solely on visual observations of the kicker during the shot without factoring in any ball trajectory data.

various aspects of soccer. Notably, some studies introduced a visual analytic system that combines video recordings with abstract visualizations of trajectory data, enabling analysts to delve deep into ball, player, or team behavior (Stein et al., 2018; Kamble et al., 2019; He, 2022). Furthermore, some comprehensive datasets have been introduced to facilitate the localization of crucial events within extended soccer video footage (Giancola et al., 2018; Deliège et al., 2021). In addition, an automatic method was proposed to localize sports fields in broadcast images, eliminating the need for manual annotation or specialized cameras (Homayounfar et al., 2017). Lastly, some analytic systems were developed to visually represent the spatiotemporal evolution of team formations, aiding analysts in understanding and tracking the dynamic aspects of soccer strategies (Wu et al., 2019; Li et al., 2023). These technological advancements have notably transformed sports analysis and enhanced the fan experience in soccer, revealing new insights and engagement opportunities. Nevertheless, unexplored possibilities persist. While previous studies have enriched our understanding of the game, untapped areas exist where technology can drive substantial advancements in soccer. For instance, incorporating predictive analytics in free-kick actions could lead to the creation of advanced algorithms that account for factors like goal distance, angle, kicker skills, defensive wall positioning, and even the goalkeeper's historical performance in stopping free kicks.

This work represents a significant step in advancing our understanding of ball-on-goal position (BoGP) in the context of free kicks directed toward the opponent's goal. Utilizing HAR backbones, we have crafted a BoGP classifier, benchmarking our models against a novel and extensive collection of free-kicks. To accomplish this, we have gathered and processed free-kick footage from various sources on the Internet. Building upon this dataset, multiple models that integrated contextual information and utilized pre-trained HAR encoders (commonly referred to as backbones) were tested to predict the final destination of the kicked ball into the goal. Notably, our methodology incorporates two crucial stages as inputs to the model: the running and the kicking stages, both depicted in Figure 1.

The significance of this approach lies in the fact that it captures the dynamic nature of a free-kick, allowing our classifier to consider the player's approach and the moment of impact. This nuanced perspective is pivotal for a more accurate and comprehensive understanding of BoGP in free kicks. Furthermore, we conducted two distinct analyses. The first analysis involved categorizing the goal into three classes (*left*, *center*, and *right*), providing a fine-grained BoGP assessment. The second analysis simplified the categorization into two classes (*left* and *right*), allowing for a broader perspective on BoGP accuracy. This dual approach enabled a deeper exploration of free-kick complexities; please refer to Figure 1.

Our contributions can be summarized as follows:

- We introduce a novel soccer free-kick dataset comprising 603 short clips from actual matches. This dataset has been curated from online sources and is readily accessible to the public.
- Through a series of experiments, we empirically

showcase the feasibility of addressing the BoGP challenge by employing a classifier that combines contextual data with a two-stream approach. Each stream offers a distinct embedding path, encompassing the running stage and the kicking stage of the free-kick process.

- Within the scope of this study, we conduct a comparative analysis of eleven different HAR backbone architectures, assessing their respective performance in BoGP classification.
- An in-depth error analysis study was undertaken to evaluate how the various classes influence the performance of the top-performing model.

The subsequent sections of this paper are structured as follows. Section 2 discusses previous related work. Section 3 outlines the proposed pipeline. Section 4 details the experimental setup and presents the results. Section 5 offers an analysis of errors. Lastly, Section 6 draws our conclusions.

2 RELATED WORK

Sports analysis has consistently captured the community's attention, leading to a substantial surge in published research over the past decade. In this sporting domain, technology has become an integral and transformative force, significantly shaping our understanding of sports, as well as how athletes train and compete. This section offers a comprehensive examination of two specific elements addressed in this study: datasets in sports and their computing application.

The available sports video datasets can be categorized into two main groups: still-image and video-sequence datasets. The first group encompasses datasets primarily designed for image classification. For instance, the UIUC Sports Event Dataset comprises 1,579 images spanning eight sports event categories (Li and Li Fei-Fei, 2007). Each category may contain subsets of images ranging from 180 to 205, categorized as easy or medium based on human subject judgments. Another noteworthy collection is the Leeds Sports Pose Dataset (Johnson and Everingham, 2010), featuring 2.000 pose-annotated images of athletes gathered from the Internet. Each image includes annotations for 14 joint locations. More recently, ultra-distance runners competitions have also been captured in wild conditions (Penate-Sanchez et al., 2020).

In contrast, the video-sequence datasets offer time series information about the actions occurring within the scene. These sequences are typically captured using stationary cameras. Sequences from individual sports provide a suitable context for activity recognition, while sequences from team sports can be used for player tracking and event detection. In this context, many sports datasets have been assembled from international competitions to advance research in automatic quality assessment for sports. Some of the most recent datasets include the MTL-AQA diving dataset (Parmar and Morris, 2019b), the UNLV AQA-7 dataset, which includes diving, gymnastic vaulting, skiing, snowboarding, and trampoline (Parmar and Morris, 2019a), and the Fis-V skating dataset (Xu et al., 2020). These datasets have been collected in controlled, non-obstructed environments, with exceptions like the UNLV AQA-7 snowboarding and skiing subsets, gathered in quiet conditions with a dark sky (night) and snowy ground.

The semantic structure of sports video content can be categorized into four layers: raw video, object, event, and semantic layers (Shih, 2018). The foundation of this pyramid consists of raw video input, from which objects are identified in the higher layers. Specifically, critical objects featured in video clips are recognized through object extraction, such as players (Guo et al., 2020) and object tracking, including the ball (Wang et al., 2019) and players (Lee et al., 2020). The event layer signifies the actions of critical objects. Various actions, combined with scene information, generate event labels that depict the related actions and interactions among multiple objects. Research in areas like action recognition (Freire-Obregón et al., 2022), re-identification (Akan and Varli, 2023; Freire-Obregón et al., 2023), facial expression recognition (Brick et al., 2018; Santana et al., 2023), trajectory prediction (Teranishi et al., 2020), and highlight detection (Gao et al., 2020) falls within the scope of this layer. The topmost layer, the semantic layer, is responsible for summarizing the semantic content of the footage (Cioppa et al., 2018). As our objective is BoGP, we seek to classify the outcome of a free-kick action. Furthermore, the mentioned collections predominantly feature professional athletes. In this context, our work does not address the team dimension, as it specifically focuses on a particular action. Nevertheless, several pivotal individuals are visible during this action, including the kicker, the referee, the other players, especially those forming the defensive wall, and the goalkeeper.

3 DESCRIPTION OF THE PROPOSAL

This paper introduces and assesses a sequential pipeline consisting of two core modules, where video

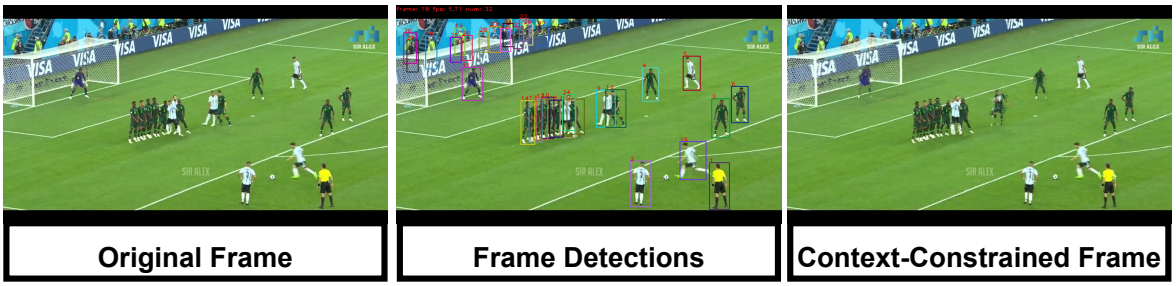

Figure 2: **Context Removal.** For every frame at time t, the process entails isolating the kicker's bounding box, which is then superimposed onto a stable background derived from the mean of τ frames.

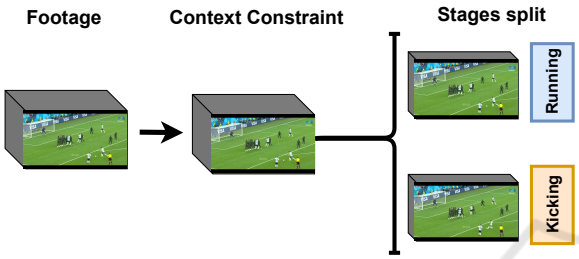

Figure 3: **Video Pre-Processing Module.** The initial video material undergoes a pre-processing phase wherein the kicker is separated from a dynamic background. Following this, two sets of frames are manually chosen to delineate the running and kicking stages. The remaining frames are excluded.

pre-processing is performed manually before entering the pipeline. The core modules include a video pre-processing module, a stage-embeddings extraction module, and a classifier. Figures 4, and 5 depict visual representations of these modules, while Figure 3 illustrates the executed video pre-processing. The following subsections comprehensively describe the video pre-processing step and each module.

3.1 Context Constraint

In order to optimize the quality of the embeddings generated by the backbone, it is imperative to ensure that the input footage provided to the action recognition networks is devoid of extraneous elements, as indicated in a prior study (Freire-Obregón et al., 2022). Within the context of the dataset utilized for the experiments detailed in this research, as described in Section 4.1, these extraneous elements encompass unrelated players, staff, supporters, and referees. Given their lack of relevance within the purpose of this work, an initial pre-processing phase is conducted to refine the raw input data by isolating the primary subject, i.e., the kicker. This task is accomplished by leveraging ByteTrack (Zhang et al., 2021), a multi-object tracking network that can precisely track the kicker within each video footage, see Figure 2. Following this, a context-constrained pre-processing technique is applied to establish an ideal setting for conducting the experiments.

In the context of acquiring context-constrained video frames for a specific kicker (k) at a given time (t) within a specified time interval ($[0,T]$), the bounding box ($BB_k(t)$) plays a crucial role. This bounding box outlines the area occupied by kicker k within the frame recorded at time t. To facilitate this process, two primary factors are considered: the bounding box area of the kicker ($BB_k(t)$) and the average number of frames required (τ) to establish a static background against which the isolated kicker (k) appears in the pre-processed video frame. The resulting pre-processed frame ($F'_k(t)$) is generated through the following equation:

$$F'_k = BB_k(t) \cup \tau$$

Here, the $\cup$ operation involves aligning and superimposing the bounding box of kicker k onto the average of the selected τ frames. This sequence of pre-processed frames constitutes the new video footage, with the kicker as the sole moving element.

Lastly, as depicted in Figure 3, the resultant footage is temporally segmented. This manual segmentation identifies two distinct moments aligned with the kicker's actions: the running stage and the kicking stage. Any elements in the video, such as the free-kick outcome or the kicker's reaction, have been excluded from the analyzed stages. This study focuses exclusively on the running stage (the phase in which the kicker approaches the ball) and the kicking stage (comprising the 16 frames before and the 16 frames after the ball is kicked).

3.2 Stage-Embeddings Extraction

The preprocessed input footage for each stage, consisting of m frames, undergoes a twofold procedure. Initially, the footage is downsampled, which results in its division into n video clips, represented as $v_1, \ldots, v_n$,

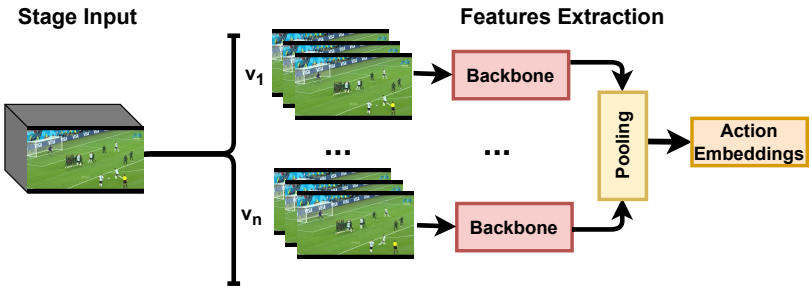

Figure 4: **Embeddings Extraction Module.** Each stage footage undergoes downsampling, dividing it into n smaller clips. A pre-trained human-action model is then applied to extract features from these clips. These features are combined using a pooling technique, resulting in a final tensor that serves as input to the classifier. This work examines two pooling methods, average pooling and max pooling.

where each clip comprises a sequence of q consecutive frames that encapsulate a snapshot of the activity, see Figure 4. In practical terms, the n clips exhibit partial overlap, spaced one frame apart from the preceding one. These video clips traverse a pre-trained HAR encoder (backbone), producing r-dimensional feature vectors. It is worth noting that these encoder models have undergone prior training on the Kinetics 400 dataset, which encompasses a broad spectrum of 400 action categories (Kay et al., 2017). Following the acquisition of feature vectors for all n video clips, a pooling layer ensures the contribution from each clip. In this regard, we have evaluated both average and max pooling layers, as seen in Section 4.

We have chosen eleven backbones to test our approach to tackle the BoGP problem. Some are more complex backbones (Slowfast or I3D) than others (the X3D instances and C2D). This section offers an overview of the HAR models considered for this study. The **C2D** (Convolutional 2D) model, designed for video action classification (Simonyan and Zisserman, 2014), exploits the power of 2D Convolutional Neural Networks (CNN) for spatial feature extraction from video frames. Its architecture comprises convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract spatial features while pooling layers reduce dimensionality to prevent overfitting. The C2D model processes each frame independently, employing CNNs to extract spatial features, which are combined to capture temporal action dynamics.

In contrast to the C2D model, the **SlowFast** model is conceived based on the principle that different video segments possess diverse temporal resolutions and contain crucial information for action recognition (Feichtenhofer et al., 2018). For example, some actions occur swiftly and necessitate high temporal resolution for detection, while others unfold more slowly and can be recognized with a lower temporal resolution. To address this variability, the SlowFast model adopts a dual-pathway approach, comprising fast and slow pathways that operate on video data at varying temporal resolutions.

Similarly, **Slow** adopts a two-stream architecture to capture both short-term and long-term temporal dynamics in videos (Feichtenhofer et al., 2021). Its slow pathway processes high-resolution frames but at a lower frame rate, similar to the C2D model. Additionally, Slow incorporates a temporal-downsampling layer to capture longer-term temporal dynamics. The Inflated 3D ConvNet (I3D) model is designed to handle short video clips as 3D spatiotemporal volumes, enabling the capture of both appearance and motion cues using a two-stream approach (Carreira and Zisserman, 2017). In this design, the first stream deals with RGB images, utilizing weights that are pre-trained on extensive image classification datasets. Simultaneously, the second stream processes optical flow images and undergoes fine-tuning in conjunction with the RGB stream.

A revised variant of the I3D model, **I3D NLN**, incorporates non-local operations to enhance spatiotemporal dependency modeling in videos (Wang et al., 2017). I3D NLN retains the two-stream architecture involving RGB and optical flow streams, processing 3D spatiotemporal volumes. In contrast to the Inception module, I3D NLN employs non-local blocks capable of learning long-range dependencies across feature map positions. By computing weighted sums of input features from all positions based on the similarity between these positions in the feature maps, I3D NLN captures global context information and improves the modeling of temporal dynamics.

Finally, we have leveraged four **X3D** model variations, distinguished by their sizes: extra small (X3D-XS), small (X3D-S), medium (X3D-M), and large (X3D-L). Each expansion incrementally transforms X2D from a compact spatial network to a spatiotemporal X3D network (Feichtenhofer, 2020) by modifying temporal (frame rate and sampling rate), spa-

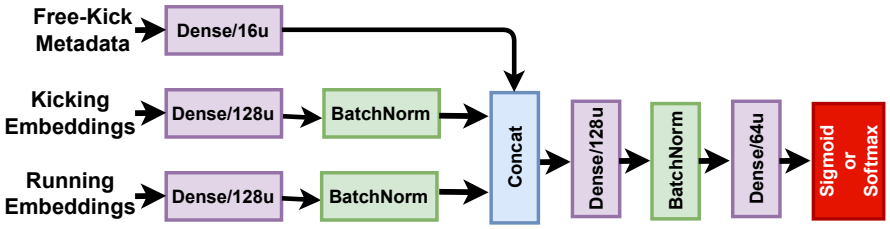

Figure 5: **The proposed classifier.** Features from the HAR backbones for running and kicking stages are processed alongside free-kick metadata, combining information from various sources to contribute to the model's decision-making process. The features extracted from the HAR backbone offer a fine-grained understanding of the kicker's movements. At the same time, free-kick metadata provides valuable context, influencing the classification outcome, particularly in diverse free-kick scenarios.

tial (footage resolution), width (network depth), and depth dimensions (number of layers and units). X3D-XS results from five expansion steps, followed by X3D-S, which includes one backward contraction step after the seventh expansion. X3D-M and X3D-L are generated by the eighth and tenth expansions, respectively. X3D-M augments the spatial resolution by elevating the spatial sampling resolution of the input video. At the same time, X3D-L expands the spatial resolution and network depth by increasing the number of layers in each residual stage.

3.3 Classifier

The proposed classifier involves feature extraction from the identical HAR backbone for both the running and kicking stages, as well as the inclusion of free-kick metadata, see Figure 5.

This three-input approach combines information from various sources, each contributing unique and complementary insights to the model's decision-making process. The features extracted from the HAR backbone offer a fine-grained understanding of the kicker's movements and actions during the free kick. Simultaneously, free-kick metadata provides valuable context and situational information that can significantly influence the classification outcome, especially when dealing with various free-kick scenarios. In this regard, the free-kick metadata encompasses four distinct input variables, each contributing specific information to the model's decision-making process. These variables include pitch side, free-kick side, free-kick distance, and kicker foot. The pitch side variable operates as a binary indicator, distinguishing between left and right. In contrast, the free-kick side variable offers a more detailed classification, representing three distinct values related to the shooting point: left to the goal, center to the goal, and right to the goal. Similarly, free-kick distance, another binary variable, provides insight into whether the free kick occurs near or far from the penalty box. Lastly, the kicker foot variable, also binary, characterizes the preferred kicking foot as either left or right.

As a result, the model receives three distinct inputs, each of which undergoes processing via dedicated fully connected layers with varying units (16 and 128) based on the nature of the input. The running and kicking paths also include batch normalization layers. Subsequently, all paths are concatenated, followed by two fully connected layers (128 and 64 units, respectively), separated by a batch normalization layer. Finally, the model's output, denoting the ball's position on the goal, is determined by either a Sigmoid or a Softmax layer, depending on whether the output comprises two or three classes.

In the conventional classification framework, the primary objective is to assign a sample to its appropriate class. In this context, we have conducted two experiments on the ball's positioning within the goal. The first experiment considers three distinct classes (*left*, *right*, and *center*), while the second experiment operates as a binary classifier, explicitly distinguishing between *left* and *right* placements. Consequently, we employ the categorical cross-entropy loss function for the first experiment:

$$Loss_1 = -\sum_{i=1}^{C} y_i \log(p_i) \tag{1}$$

Where C is the number of classes, y_i is the true probability distribution (one-hot encoded vector) of the ground truth class, and p_i is the predicted probability for class i. For the second experiment, the considered loss function to tackle the problem is the binary cross-entropy:

$$Loss_2 = \frac{-1}{N} \sum_{i=1}^{N} -(y_i \log(p_i) + (1-y_i) \log(1-p_i)) \tag{2}$$

Where p_i is the i-th scalar value in the model output, y_i is the corresponding target value, and N is the number of scalar values in the model output.

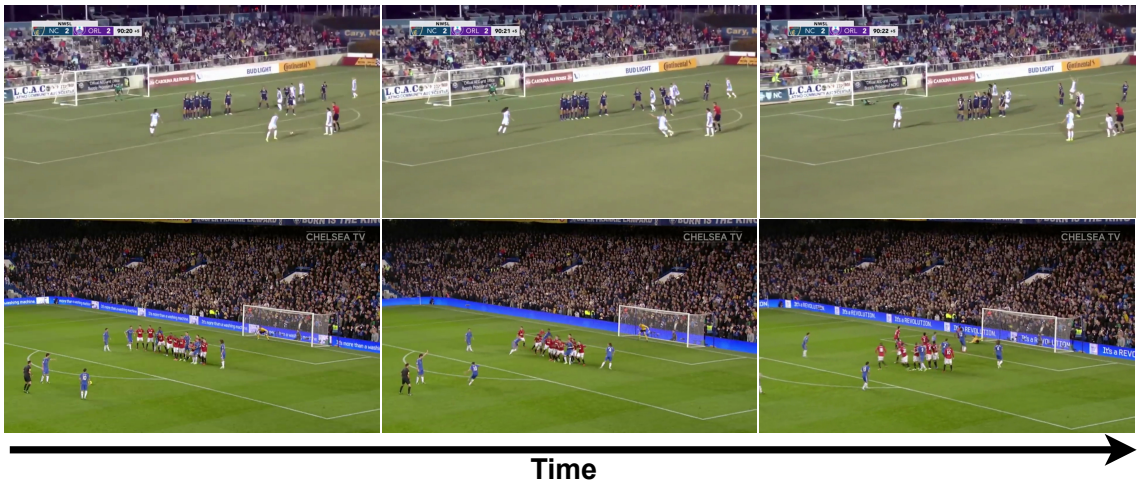

Figure 6: **Free-kick Dataset Sequences.** The video dataset used in this study was gathered from the Internet without any imposed usage restrictions. Due to this unrestricted collection approach, the dataset exhibits notable pose, scale, and lighting conditions variability. Video clips were carefully edited to retain frames from just before the running stage until the moment of the free-kick outcome.

4 EXPERIMENTAL SETUP

This section is divided into three subsections related to the dataset acquisition, experimental setup, and achieved results of the designed experiments. The first subsection provides technical details regarding the dataset, including its acquisition and data-cleaning processes. The second subsection outlines the technical aspects of our proposal, such as the data split. Finally, the third subsection summarizes the achieved results.

4.1 Dataset

To our knowledge, there is no publicly available soccer free-kick dataset. Our data collection approach hinges on generality, intending to construct robust detection models for practical use. This compilation of videos was sourced from the Internet without any usage restrictions, resulting in considerable variations in pose, scale, and lighting conditions, see Figure 6. The data collection process encompasses three steps:

1. **Web Scraping:** an extensive search was conducted to gather relevant images using keywords like ”free-kick soccer,” ”free-kick compilation”, and the names of various soccer players well known for frequently shooting free kicks.
2. **Shot Labeling:** labeling involves carefully editing each video clip. These clips are trimmed to cover the period from just before the kicker initiates the run to the occurrence of the shot outcome. This stage results in a subset of 603 free-kick clips.
3. **Manual Annotation:** each free-kick clip is manually reviewed and annotated. Annotations encompass various variables, including pitch side, free-kick side, free-kick distance, kicker foot (left or right), kick outcome, barrier configuration, gender, goalkeeper zone, and the specific frame in which the ball is kicked. The resolution of these clips is 1920×1080 pixels.

Despite the initial inclusion of 603 free-kick clips in the dataset, several factors reduced this number. A critical consideration was the camera viewpoint, as it played a substantial role in the selection process. To maintain shooting action stability, clips where the camera perspective was positioned behind the goalkeeper or the kicker were excluded. As described in Section 3.1, the remaining 584 videos underwent people detection using ByteTrack. Unfortunately, some videos exhibited low image quality, resulting in subpar detection performance. As a consequence, the dataset was further reduced to 539 clips.

Subsequently, the duration of the videos became a focal point, as clips that were excessively short in length were found to be inadequate for extracting meaningful information. For instance, videos commencing precisely as the player initiated the kick (without a preceding running stage) were omitted from consideration due to the need for a minimum frame count to extract pertinent information. All clips containing fewer than 32 frames were accordingly excluded, ultimately reducing the dataset to 451 clips.

Table 1: **Comparative Performance Analysis of HAR Architectures for BoGP Estimation when Considering Three Classes**. This table compares different backbone architectures used to detect BoGP during free-kick shots. The first column lists the backbone models, while the second column specifies the number of frames the model utilizes for generating HAR embeddings. The table includes crucial performance metrics such as the number of frames per embedding backbone, the applied pooling method, and the values of the performance metrics: accuracy, precision, recall, and F1-Score.

Backbone	#Frames	Pooling	Accuracy	Precision	Recall	F1-Score
C2D (Simonyan and Zisserman, 2014)	8	Average	52.9%	49.4%	43.1%	46.1%
I3D (Carreira and Zisserman, 2017)	8	Average	51.4%	42.7%	39.6%	41.1%
I3D NLN (Wang et al., 2017)	8	Average	51.9%	44.6%	41.2%	42.8%
Slow4x16 (Feichtenhofer et al., 2021)	4	Average	55.0%	49.4%	44.6%	46.9%
Slow8x8 (Feichtenhofer et al., 2021)	8	Average	55.3%	46.1%	41.5%	43.7%
SlowFast4x16 (Feichtenhofer et al., 2018)	32	Max	55.0%	47.1%	43.9%	45.4%
SlowFast8x8 (Feichtenhofer et al., 2018)	32	Average	53.4%	47.4%	45.2%	46.2%
X3D-XS (Feichtenhofer, 2020)	4	Max	51.2%	46.3%	43.9%	45.1%
X3D-S (Feichtenhofer, 2020)	4	Max	53.4%	44.9%	43.5%	44.2%
X3D-M (Feichtenhofer, 2020)	13	Max	53.6%	47.9%	43.0%	45.3%
X3D-L (Feichtenhofer, 2020)	16	Average	**57.2%**	**50.0%**	**48.5%**	**49.3%**

The problem's intrinsic nature also emerged as a significant determining factor during clip selection. Specifically, any clips in which the kick did not successfully reach the goal, such as instances where the ball failed to surpass the defensive barrier, were omitted. In such cases, it was infeasible to ascertain the target location within the goal, rendering these clips inapplicable. Therefore, a refined subset of 418 clips was designated for inclusion in this study.

4.2 Experimental Setup

The results presented in this section refer to the average accuracy on five repetitions of 10-fold cross-validation for each experiment. Significantly, the class distribution within the dataset is characterized as follows: 187 free-kick shots are directed towards the left side of the goal, 181 are aimed at the right side, and 50 target the center area of the goal. The class distribution exhibits a notable imbalance, particularly in the case of the center-side shots. We have implemented a class weighting strategy during the model training phase to address this issue. The adjustment of class weights in the training process serves to amplify the model's sensitivity to minority classes, effectively mitigating the inherent challenge of disparate class distributions. This approach serves as a valuable mechanism to rectify any potential bias arising from the overrepresentation of majority classes, thereby ensuring equitable model performance across all classes.

4.3 Results

Table 1 presents a comparative performance analysis of various HAR backbone architectures utilized to estimate the BoGP during free-kick shots, specifically when considering three different target classes: *left*, *center*, and *right*. The table highlights the number of frames used for each embedding backbone (denoted as q in Section 3.2), the pooling method employed, and key performance metrics including accuracy, precision, recall, and F1-Score. The presented HAR backbone architectures encompass a range of models described in Section 3.2. Each model is evaluated based on the aforementioned metrics, providing valuable insights into their effectiveness in BoGP estimation during free-kick situations.

A noteworthy observation pertains to the choice of pooling layers for the HAR embeddings (see Figure 4). The data presented in Table 1 reveals an intriguing trend: lighter models, exemplified by the X3D instances, tend to favor the utilization of the MaxPool layer, while heavier models typically demonstrate a preference for the AveragePool layer. This distinction in pooling layer selection reflects these models' diverse architectural considerations and requirements, underscoring the need to suit the pooling method to the specific characteristics and demands of a given HAR model.

The table prominently illustrates the distinct performance levels exhibited by various models. X3D-L, in particular, stands out as the top performer, boasting the highest accuracy (57.2%), precision (50.0%), recall (48.5%), and F1-Score (49.3%). Following closely in classification performance are the SlowFast and Slow instances, although they lag by a margin of

Table 2: **Comparative Performance Analysis of HAR Architectures for Soccer Player Free-Kick Shoot Zone Estimation when Considering Two Classes**. This table compares different backbone architectures used to detect soccer player shoot zones during free-kick shots. The first column lists the backbone models, while the second column specifies the number of frames the model utilizes for generating HAR embeddings. The table includes crucial performance metrics such as the number of frames utilized, the pooling method applied, accuracy, precision, recall, and F1-Score. These metrics offer valuable insights into the effectiveness of each backbone architecture for this specific task.

Backbone	#Frames	Pooling	Accuracy	Precision	Recall	F1-Score
C2D (Simonyan and Zisserman, 2014)	8	Max	67.4%	56.5%	60.2%	58.3%
I3D (Carreira and Zisserman, 2017)	8	Average	63.1%	51.3%	56.4%	53.7%
I3D NLN (Wang et al., 2017)	8	Max	62.8%	52.6%	51.7%	52.2%
Slow4x16 (Feichtenhofer et al., 2021)	4	Average	66.9%	**60.0%**	68.3%	63.9%
Slow8x8 (Feichtenhofer et al., 2021)	8	Max	65.8%	57.6%	67.7%	62.2%
SlowFast4x16 (Feichtenhofer et al., 2018)	32	Average	**69.1%**	57.7%	**76.1%**	**65.7%**
SlowFast8x8 (Feichtenhofer et al., 2018)	32	Max	63.6%	56.6%	69.9%	62.5%
X3D-XS (Feichtenhofer, 2020)	4	Max	61.9%	47.2%	48.3%	47.7%
X3D-S (Feichtenhofer, 2020)	4	Average	64.4%	50.9%	74.3%	60.4%
X3D-M (Feichtenhofer, 2020)	13	Max	66.0%	58.4%	51.4%	54.7%
X3D-L (Feichtenhofer, 2020)	16	Max	65.8%	59.7%	56.6%	58.1%

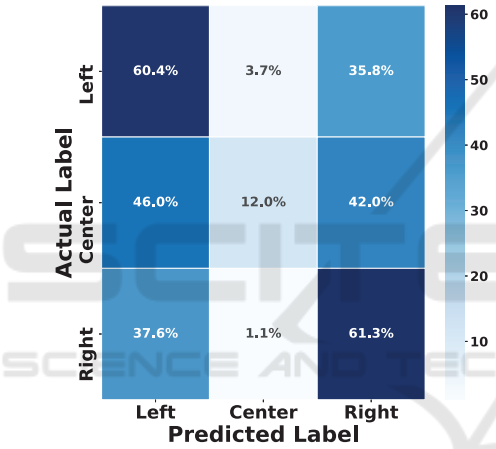

Figure 7: Three-class SlowFast4x16 confusion matrix.

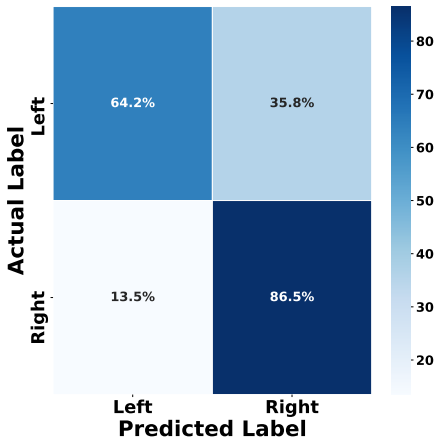

Figure 8: Two-class SlowFast4x16 confusion matrix.

2.2% in accuracy. It is worth noting that the overall performance in the context of three-class classification remains relatively modest, as evidenced by the F1-Score, though exceeding that of a random classifier. Section 5 provides a comprehensive error analysis.

To complete our evaluation, Table 2 presents a comparative performance analysis of various HAR backbone architectures used in soccer player free-kick shoot zone estimation when considering two classes: *left* and *right*. Once again, the models are evaluated in this scenario based on details about the number of frames used, the pooling method applied, and the four key performance metrics: accuracy, precision, recall, and F1-Score. Comparing this table with the previously discussed Table 1, we observe an interesting transition regarding the number of classes considered. The simplification of the classification task has a notable impact on model performance. Despite the reduced complexity of the classification problem, there are variations in the performance of the backbone architectures, indicating that the choice of backbone remains critical. Performance-wise, several observations can be made. For instance, SlowFast4x16 exhibits the highest accuracy (69.1%) in this two-class classification scenario, outperforming other models. Additionally, Slow4x16 achieves a remarkable 60.0% precision, indicating its ability to accurately classify instances. The F1-Score, which combines precision and recall, is also noteworthy, with SlowFast4x16 leading the way with a score of 65.7%. These metrics provide valuable insights into the effectiveness of the backbone architectures for the specific task of soccer player free-kick shoot zone estimation. In contrast to the outcomes in the three-class scenario, the utilization of MaxPool and AveragePool layers is evenly distributed in this table.

The architecture of the classifier described in Section 3.3 poses an intriguing question: how does the

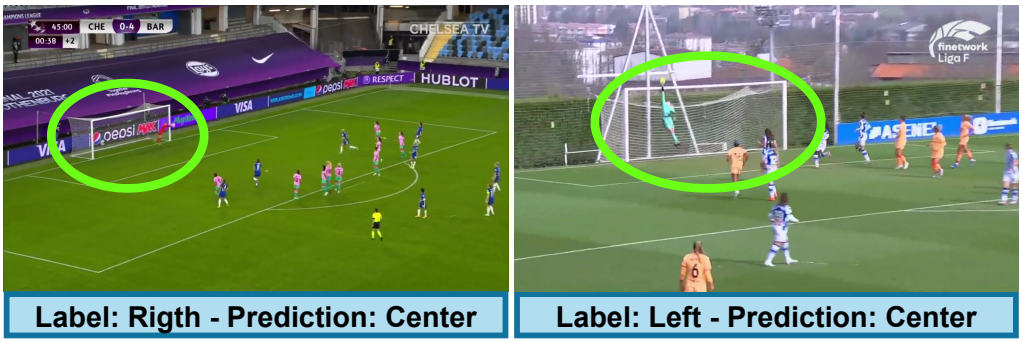

Figure 9: **SlowFast4x16 Misclassified Clips.** These frames represent the ultimate phase of two distinct samples. It is important to note that the proposed model exclusively examines the actions of the kicker, meaning it does not consider any frames beyond the 16 post-kicking frames, and the background remains static. Consequently, the frames presented in this figure were never seen by the models; they are included solely to exemplify the intricacies associated with the *center* class. Notably, the classifier erroneously categorizes these clips as *center* when labeled as *right* and *left*, respectively.

incorporation of free-kick metadata impact performance? Upon calculating the mean accuracy across all scrutinized models, the obtained outcome indicates that without consideration for free-kick metadata, the accuracy diminishes by 3%, and the F1-Score experiences a 4% decline. This signifies that metadata enhances contextual information regarding free-kick embeddings, yet it does not introduce bias to the proposed model.

In summary, as shown in this table, the transition from a three-class to a two-class problem emphasizes the consequences of simplifying the classification task. It underscores the performance differences among various HAR backbone architectures and their potential suitability for specific sports action recognition tasks. However, these findings have raised several questions, including the influence of the center class on classification, the distribution of error predictions, and the examination of confusion matrices for the top-performing models. These questions will be addressed in the following section.

5 ERROR ANALYSIS

In this section, our primary focus is on the top-performing model, which employs the SlowFast4x16 backbone. It is crucial to comprehensively analyze its performance under scenarios involving two and three classes. As a case in point, Figures 7 and 8 visually represent the confusion matrices for both experimental settings.

Our analysis presents the confusion matrix for our classification model, designed to categorize free-kick soccer actions into one of three classes: *left*, *center*, or *right*. As illustrated in Figure 7, this matrix provides valuable insight into the model's performance and ability to classify BoGP correctly. The diagonal elements of the matrix represent instances where the model's predictions align with the actual classes. For instance, the model achieved an accuracy of approximately 60.4% in identifying *left* shots, 12.0% for *center*, and 61.3% for *right*. These values indicate the model's proficiency in correctly classifying shots into their respective categories. However, the off-diagonal elements reveal cases of misclassification. Notably, there is some confusion between the *center* and the other two classes. The model often misclassifies *center* shots as *left* (46.0%) or *right* (42.0%), suggesting improvement in distinguishing *center* shots from the others. Additionally, *left* shots are occasionally misclassified as *right* (35.8%), and *right* shots are occasionally mislabeled as *left* (37.6%) or *center* (1.1%).

Our analysis suggests that the classifier faces challenges in accurately distinguishing the *center* category, as illustrated in Figure 9. The intricacies of this classification become apparent, even for human annotators, as the camera perspective can sometimes obscure the goal's position. This issue is compounded by the limited number of *center* samples, coupled with the wide range of camera angles in the dataset. Consequently, achieving a fine-grained classification for *center* may not be practically feasible given these constraints.

The confusion matrix shown in Figure 8 suggests a notable accuracy in classifying instances, particularly on the diagonal elements. The top-left quadrant indicates a correct classification rate of 64.2% for the *left* category, while the bottom-right quadrant signifies an 86.5% accuracy in classifying the *right* category. However, there is some misclassification, as evidenced by the off-diagonal elements, with 35.8% of *left* instances being erroneously classified as *right* and 13.5% of *right* instances being misclassified as *left*.

6 CONCLUSIONS

In conclusion, this study extensively examined the performance of various HAR backbone architectures in estimating the BoGP during free-kick shots. The investigation covered three-class (*left*, *center*, and *right*) and two-class (*left* and *right*) classification scenarios, providing valuable insights into the effectiveness of different models.

X3D-L emerged as the top performer for the three-class classification with notable accuracy, precision, recall, and F1-Score. However, the overall performance in this context remained modest, prompting a comprehensive error analysis in Section 5. In contrast, the two-class scenario revealed a transition in the number of classes and demonstrated that despite the reduced complexity, the choice of backbone architecture remains critical. SlowFast4x16 exhibited the highest accuracy and noteworthy precision and F1-Score, highlighting its effectiveness in soccer player free-kick shoot zone estimation. The inclusion of Free-kick metadata in the analysis showcased its impact on performance, revealing a 3% accuracy drop and a 4% decline in F1-Score when not considered. Importantly, this decline signifies the role of metadata in enhancing contextual information without introducing bias to the model.

The focus on the top-performing model, SlowFast4x16, included a detailed examination of confusion matrices for the three-class and the two-class scenarios. While the model demonstrated proficiency in classifying instances, challenges were identified, particularly in distinguishing the *center* category. The limited number of samples and diverse camera angles posed practical challenges in achieving fine-grained classification for *center*.

These findings highlight the complexity of sports action recognition tasks, emphasizing the need for careful consideration of the model architecture and task simplification's influence. Further questions were raised, including the impact of the center class on classification, the distribution of error predictions, and the exploration of confusion matrices for top-performing models, providing avenues for future research and improvement.

ACKNOWLEDGEMENTS

This work is partially funded by the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22 and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04.

REFERENCES

Akan, S. and Varli, S. (2023). Reidentifying soccer players in broadcast videos using body feature alignment based on pose. In *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, page 440–444, New York, NY, USA. Association for Computing Machinery.

Brick, N. E., McElhinney, M. J., and Metcalfe, R. S. (2018). The effects of facial expression and relaxation cues on movement economy, physiological, and perceptual responses during running. *Psychology of Sport and Exercise*, 34:20–28.

Carreira, J. and Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Cioppa, A., Deliège, A., and Van Droogenbroeck, M. (2018). A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1846–184609.

Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Droogenbroeck, M. V. (2021). Soccernet-v2 : A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Deloitte (2023). Deloitte football money league 2023. Accessed on November 3, 2023.

Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2018). Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R. B., and He, K. (2021). A large-scale study on unsupervised spatiotemporal representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308.

FIFA (2022). 2019-2022 revenue. Accessed on November 3, 2023.

Freire-Obregón, D., Lorenzo-Navarro, J., Santana, O. J., Hernández-Sosa, D., and Castrillón-Santana, M. (2022). Towards cumulative race time regression in sports: I3D ConvNet transfer learning in ultra-distance running events. In *International Conference on Pattern Recognition (ICPR)*, pages 805–811.

Freire-Obregón, D., Lorenzo-Navarro, J., Santana, O. J., Hernández-Sosa, D., and Castrillón-Santana, M. (2023). A Large-Scale Re-identification Analysis in Sporting Scenarios: the Betrayal of Reaching a Critical Point. In *International Joint Conference on Biometrics (IJCB)*.

Gao, X., Liu, X., Yang, T., Deng, G., Peng, H., Zhang, Q., Li, H., and Liu, J. (2020). Automatic key moment extraction and highlights generation based on comprehensive soccer video understanding. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6.

Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. (2018). Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–1810.

Guo, T., Tao, K., Hu, Q., and Shen, Y. (2020). Detection of ice hockey players and teams via a two-phase cascaded cnn model. *IEEE Access*, 8:195062–195073.

He, X. (2022). Application of deep learning in video target tracking of soccer players. *Soft Computing*, 26(20):10971–10979.

Homayounfar, N., Fidler, S., and Urtasun, R. (2017). Sports field localization via deep structured models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4012–4020.

Johnson, S. and Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*, pages 12.1–11.

Kamble, P., Keskar, A., and Bhurchandi, K. (2019). A deep learning ball tracking system in soccer videos. *Opto-Electronics Review*, 27(1):58–69.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The Kinetics Human Action Video Dataset. *CoRR*.

Lee, J., Moon, S., Nam, D.-W., Lee, J., Oh, A. R., and Yoo, W. (2020). A study on sports player tracking based on video using deep learning. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1161–1163.

Li, L. and Li Fei-Fei (2007). What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Li, L., Zhang, T., Kang, Z., and Zhang, W.-H. (2023). Design and implementation of a soccer ball detection system with multiple cameras. *ArXiv*, abs/2302.00123.

Microsoft (2023). Shaping the future of the game. Accessed on November 3, 2023.

Parmar, P. and Morris, B. (2019a). Action quality assessment across multiple actions. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1468–1476. IEEE.

Parmar, P. and Morris, B. T. (2019b). What and how well you performed? A multitask learning approach to action quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 304–313. Computer Vision Foundation / IEEE.

Penate-Sanchez, A., Freire-Obregón, D., Lorenzo-Melián, A., Lorenzo-Navarro, J., and Castrillón-Santana, M. (2020). TGC20ReId: A dataset for sport event re-identification in the wild. *Pattern Recognition Letters*, 138:355–361.

Santana, O. J., Freire-Obregón, D., Hernández-Sosa, D., Lorenzo-Navarro, J., Sánchez-Nielsen, E., and Castrillón-Santana, M. (2023). Facial expression analysis in a wild sporting environment. *Multimedia Tools and Applications*, 82(8):11395–11415.

Shih, H.-C. (2018). A survey of content-aware video analysis for sports. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5):1212–1231.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *ArXiv*, abs/1406.2199.

Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, G., Grossniklaus, M., and Keim, D. A. (2018). Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):13–22.

Teranishi, M., Fujii, K., and Takeda, K. (2020). Trajectory prediction with imitation learning reflecting defensive evaluation in team sports. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 124–125.

Wang, S., Xu, Y., Zheng, Y., Zhu, M., Yao, H., and Xiao, Z. (2019). Tracking a golf ball with high-speed stereo vision system. *IEEE Transactions on Instrumentation and Measurement*, 68(8):2742–2754.

Wang, X., Girshick, R. B., Gupta, A. K., and He, K. (2017). Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803.

Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S., and Chen, W. (2019). Forvizor: Visualizing spatio-temporal team formations in soccer. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):65–75.

Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.-G., and Xue, X. (2020). Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2021). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *European Conference on Computer Vision*.