Intelligent Soccer Event Detection and Highlights Generation with Broadcast Cues Integration

Anirudh Narayanan[†], Sergei Chuprov[*], Leon Reznik[†], Raman Zatsarenko[†], and Dmitrii Korobeinikov[†]
[*]The University of Texas Rio Grande Valley, Edinburg, TX, USA,
Email: sergei.chuprov@utrgv.edu
[†]Rochester Institute of Technology, Rochester, NY, USA,
Email: an9425@g.rit.edu, leon.reznik@rit.edu, rz4983@rit.edu, dk9148@rit.edu

***Abstract*—In this paper, we present an innovative approach to automate key event detection and highlights generation from soccer match videostreams that allows to improve accuracy and reliability, as well as to reduce data consumption and training time. Our method segments the videostream into distinct frames based on camera angles and activities, and integrates intelligent video analytics with additional visual information provided by broadcasters. As our major novelty in comparison to other intelligent soccer video analysis approaches, we deploy a Multi-Class Image Classifier to segment the video into wide-angle overviews, close-ups, and in-game replays, which allows us to improve the event detection performance and the quality of generated highlights. As another major novelty, we leverage YOLOv8 to detect events such as bookings, substitutions, and goals based on the additional information portrayed by broadcasters during the game. We evaluate our approach and demonstrate its advantage, when the additional information from broadcasters is available, against existing ones that analyze only the actions happening in the scene itself, such as the players' current positions and their actions between the frames. We evaluate our pipeline on a real soccer game recording and compare the highlights it generates with the official highlights provided by the broadcaster. Our pipeline demonstrates ample performance and is able to detect all key events in the game.**

***Index Terms*—Soccer event detection, soccer highlights generation, intelligent video analysis**

I. Introduction

Soccer, also known as football, captivates billions of spectators globally. However, some soccer enthusiasts may find the duration of a particular match too lengthy for their viewing preferences and choose to rely on game highlights produced by such entities as TV broadcasters and soccer events organizers. These broadcasts typically require manual effort and expertise, which presents an opportunity for developing intelligent tools capable of analyzing game footage and generating engaging highlights by processing in-game events. Such tool could revolutionize content creation, offering broadcasters significant time and cost savings by automating the highlight generation process.

Event detection and highlights generation from soccer matches is a challenging task, primarily due to the dynamic nature of the game and varying camera angles in the broadcast. The research community has extensively explored diverse methodologies to identify key events for statistical analysis and automated highlights generation. Among the Machine Learning (ML) technologies employed, YOLO object detection models [9] have played a significant role in numerous projects. Darapaneni *et al.* [2] evaluate the effectiveness of Faster-RCNN models utilizing VGG16 and ResNet50 architectures alongside YOLO. Sha *et al.* [10] also employ YOLO to pinpoint players on the field and deduce the homography matrix from the camera feed and sport-specific details. Cioppa *et al.* [1] follow the approach developed in [10] and demonstrate a Camera Calibration for Broadcast Videos framework built on a three-stage process, each facilitated by a distinct neural network. Integrating the available data of various modalities is another effective event detection way studied by Raventos *et al.* [8]. The authors introduce an innovative approach of integrating both audio and visual descriptors to analyze and extract events from soccer games. This method stands out from traditional models that primarily focus on analyzing visual information, as employing audio cues enables adding a second layer of validation that enhances the reliability of event detection. However, the fusion of audio and video flows for information extraction is another challenge that needs to be addressed in this case.

The usage of additional data modalities and available information provided by the broadcasters in their soccer streams for improving intelligent event detection and highlights extraction is still under-researched. In this paper, we introduce a novel approach to automate key events detection and highlights generation from soccer matches videostreams by: **(1)** segmenting the videostream into distinct frames corresponding to various camera angles and activities (repeat of the in-game event or actual game); and **(2)** fusing the intelligent video analytics with the processing of additional visual information added by broadcasters to their streams. The dynamic broadcast nature of soccer, characterized by constant camera transitions to capture the most engaging angles, presents a significant challenge for events tracking and detection. To mitigate this, we deploy a multi-class image classifier that segments the videostream into three distinct parts: segments containing frames covering the wide-angle overviews of the pitch; segments containing frames covering the close-ups of the players, referees, coaches, and audience reactions; and segments representing the in-game replays of key moments as selected by the broadcasters.

This research was partially supported by the USA National Science Foundation (award # 2321652)

Our results demonstrate that dividing the broadcast video into these segments prior to employing a video classification model significantly enhances its accuracy. This method allows the event detection and classification model to more confidently learn and interpret the temporal changes occurring between frames, leading to more precise event detection.

In many soccer leagues, broadcasters enhance live feeds with real-time additional information, which may include the scoreboard demonstrated in some part of the stream; the transcription of the events into the text portrayed in specified blocks, such as yellow or red cards assigned to a player; or a transition between the actual game and the repetition of the last key event in the game. We integrate YOLOv8 into our events detection and highlights generation pipeline to exploit these features, offering a significant enhancement over conventional soccer games event detection approaches that typically overlook such broadcast elements and analyze only the players' positions and actions. We provided an open access to our implementation at the GitHub repository[1].

II. Developed Events Classification and Highlights Generation Pipeline

We develop a software pipeline consisting of multiple modules, each responsible for specific tasks, working together to enhance the performance of key event detection and highlights generation from soccer matches. The initial soccer match videostream, typically 90 minutes long, is processed to identify key events using three distinct ML models: YOLOv8, Multi-Class Image Classifier, and Video Masked Autoencoder (VideoMAE). The first module, the Additional Information Extraction module, employs YOLOv8 to detect specific events such as bookings (yellow and red cards), substitutions, and scoreboards portrayed by the broadcaster during the game. Every 10th frame from the input video is analyzed by YOLOv8 to identify and track these events. Regardless of whether an event is detected, the scoreboard is consistently monitored and sent to the scoreboard tracker.

The second module, the Scoreboard Tracker, is a computer vision script that employs OpenCV[2] library. This module utilizes a Region of Interest (ROI) identified by the YOLOv8 model, extracts it and sends to Amazon Rekognition[3], which helps to keep track of goals in the game. Since the scoreboard is updated almost instantaneously, this approach ensures an accurate track of the game's score. The employment of Rekognition API provided by Amazon allows us to maintain reliable and highly accurate score extraction without deploying and training in-house optical character recognition models. By integrating this module into our pipeline, we maintain a real-time and precise record of the score throughout the match.

The third module, the Multi-Class Image Classification module, is used to differentiate between the live game footage and in-game replays, and further segment the game into Close-Up and Overview frames. Broadcasters often replay key events during the game, indicated by a transition frame of the league logo, such one can see in Figure 2(e). The classifier identifies these transitions and divides the footage into the actual game segments and replays. After segmenting, the actual game frames are further split into Overview and Close-Up frames, which are processed in a separate manner. As our results show, this approach increases the accuracy of the VideoMAE events detection and classification. The splitting is beneficial due to the dynamic way broadcasters present the game, which may confuse the model and lead to misclassifications.

The final VideoMAE Events Detection and Classification module processes the overview frames to detect events such as Penalty Kicks and Set-Pieces (free-kicks and corners). By splitting the gameplay into the Overview and Close-Up frames, the performance of the VideoMAE events detection is improved, reducing false positives and enhancing the overall reliability. Our results, presented in sec. III, demonstrate that this module is crucial for identifying key moments in the game that significantly impact the quality of the highlights produced.

After processing by all modules is done, a specified Python script is utilized to compile all extracted events and replays into a single highlights video. The script combines the extracted events based on their timestamps (in the ascending manner), and compares them between extracted events and replays to avoid duplicates.

In summary, our architecture is based on a synergy between multiple interconnected modules to improve the performance of key event detection and highlights generation from soccer matches. In Figure 1, we depict our architecture, represent modules, and show interconnections between them. Below we describe the process of our pipeline development, and the technicalities of each component in more detail.

A. Dataset

To train and verify our pipeline, we utilized the open-source SoccerNetV2 dataset presented by Giancola *et al.* [3], which is designed to foster contributions towards the automation of soccer analytics. The dataset consists of video footage of 550 games from the top football leagues in the world. The dataset offers video footage in two resolutions: low definition at 240p, employed in our paper, and high definition at 1024p at 25 frames per second. The dataset offers 550 games for both resolutions. Figure 2 demonstrates some examples of the frames from the dataset.

B. Employed Models

To integrate events detection module in our pipeline, we initially tested three various ML architectures: Convolutional Long-Short Term Memory (LSTM) model for classifying sequential data, Long-term Recurrent Convolutional Network (LRCN) for handling sequential video data, and VideoMAE [4]. We employed the YOLOv8 model to detect certain events in the broadcast video, such as the actual score changes and substitutions. To further improve the soccer events detection performance in the video footage, we employed and further re-trained the Multi-Class Image Classifier, initially proposed

[1] https://github.com/Anirudhrn98/Event_Detection_Soccer

[2] https://opencv.org/

[3] https://aws.amazon.com/rekognition/

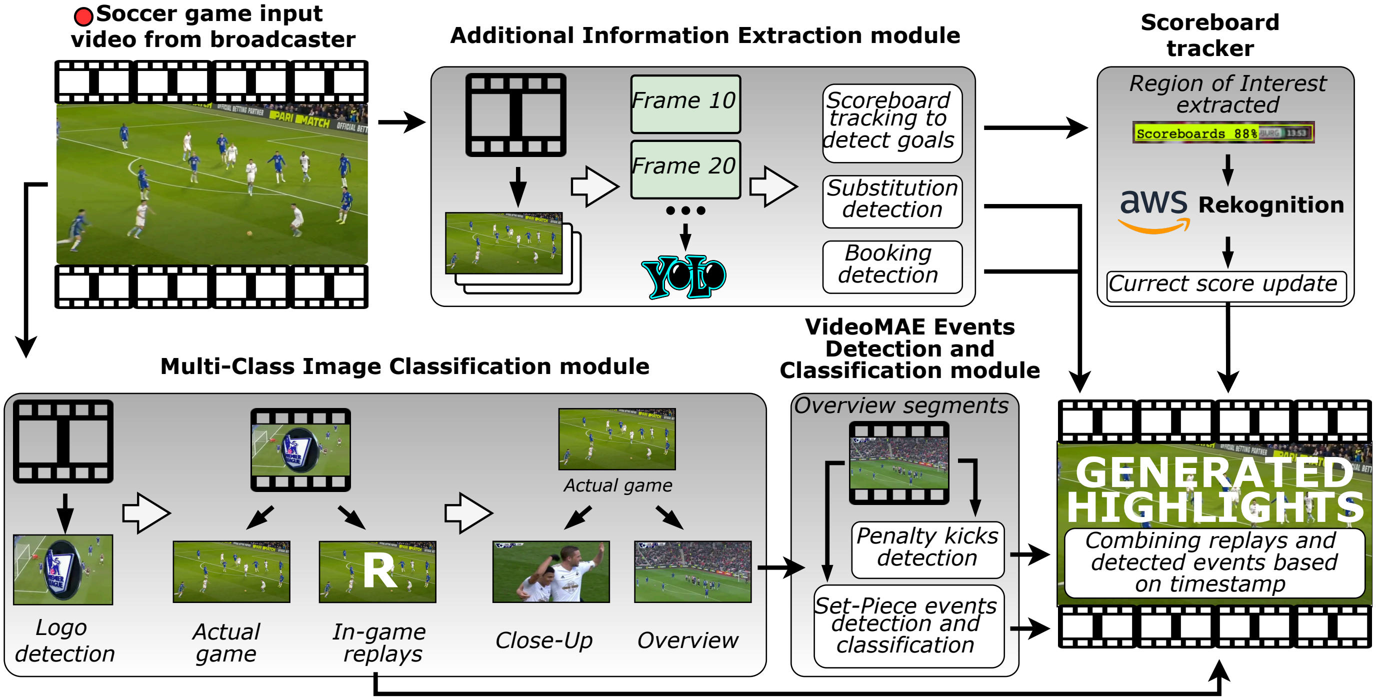
Soccer game input
video from broadcaster
Additional Information Extraction module
Frame 10
Frame 20
YOLO
Scoreboard tracking to detect goals
Substitution detection
Booking detection
Scoreboard tracker
Region of Interest extracted
Scoreboards 88%
aws Rekognition
Currect score update
Multi-Class Image Classification module
Logo detection
Actual game
In-game replays
Actual game
Close-Up
Overview
VideoMAE Events Detection and Classification module
Overview segments
Penalty kicks detection
Set-Piece events detection and classification
GENERATED HIGHLIGHTS
Combining replays and detected events based on timestamp

Fig. 1. Architecture of the developed soccer events detection and highlights generation pipeline. The pipeline consists of the following modules: Additional Information Extraction module, which employs YOLOv8 for detecting events like Booking and Substitution from the additional information portrayed by broadcasters; Scoreboard tracker that leverages Amazon Rekognition to track goals from the extracted scoreboard information by YOLO; Multi-Class Image Classifier for segmenting live footage and replays showed by broadcasters during the game, and further diving the live footage into Close-Up and Overview segments; and VideoMAE for detecting Penalty Kicks and Set-Piece events from the Overview segments

in [6], to differentiate between various camera angles and repeats shown by broadcasters during the game. In addition to the open-source models mentioned above, we employed Rekognition ML service provided by Amazon to extract the additional visual information added by broadcasters to their streams.

C. Initial Training and Model Comparison

Our initial approach trained models to identify key events from the dataset, but they struggled to differentiate between Free-Kicks and Corners due to visual similarities like player clustering and motion near the penalty box. To improve detection, we combined these into a Set-Piece class, as their distinction had minimal impact on game highlights. We also introduced an Open-Play category for segments without specified events. The models were trained on a modified dataset with 9 event categories: Transition, Set-Piece, Offside, Card, Goals, Penalty, Open-Play, Celebration, and Substitution, each with an average of 150 videos ranging from 3-10 seconds. We used Accuracy to evaluate performance. The LSTM model was trained for 100 epochs with early stopping at 10 epochs of no improvement in validation loss, which plateaued around the 7th epoch. Despite using dropout layers and L2 regularization, no significant improvement in generalization was observed. Results are in Table I.

TABLE I
COMPARISON OF LSTM, LRCN, AND VIDEOMAE MODELS ACCURACY (%) FOR DETECTING SPECIFIED EVENTS IN SOCCER GAME

	Model		
Event class	**LSTM**	**LRCN**	**VideoMAE**
Transition	83.3	85.2	100
Set-Piece	21	57.1	94.64
Card	66.6	66.6	88.89
Goals	76.5	88.2	82.61
Penalty	85.9	92.9	95.83
Open-Play	70	60	97.37
Substitution	45.1	77.4	81.82
Celebration	53.7	72.2	83.33
Average	72.86	82.5	92

The LRCN model was trained for 200 epochs, including an early stopping mechanism with a patience of 15 epochs on the validation loss. The Accuracy of both the LSTM and the LRCN models for each class is shown in Table I. As one can see, the LRCN model performed better on the test data and was able to detect soccer match events more accurately. However there are certain classes such as Open-Play and Set-Piece, where even the LRCN model failed to show adequate results. In pursuit of a more accurate model for event detection within a soccer game, we decided to employ a VideoMAE model that was pre-trained on an extensive dataset [4].

The VideoMAE re-training included fine-tuning of the model pre-trained on large video classification dataset [4]. The initial re-training across nine classes yielded the average Accuracy better than both LSTM and LRCN (see Table I). Nonetheless, when applied to a three-minute video segment,

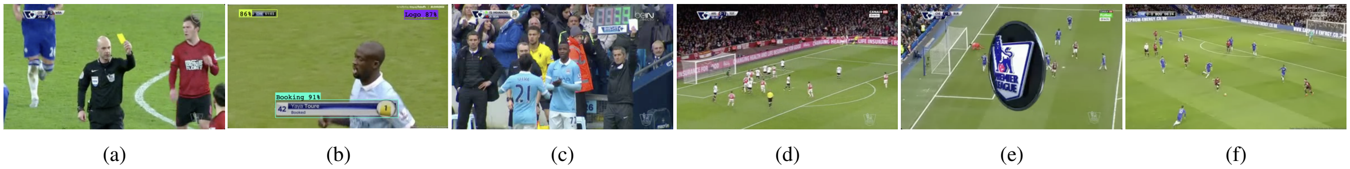
(a) (b) (c) (d) (e) (f)

Fig. 2. Examples of various frames from the employed dataset: (a) – Booking (card showed to a player by a referee); (b) – additional visual information related to the Booking event portrayed by the broadcaster and detected by YOLO; (c) – Substitution, Close-Up scene; (d) – example of a Set-Piece event (corner), Overview scene; (e) – Transition event (League Logo demonstrated by broadcasters during the replays); (f) – Open-Play event, Overview scene

the model produced numerous misclassifications. Further analysis suggested that discrepancies between the homogeneity of training data and the heterogeneity of actual game footage contributed to this issue. Training samples were sourced mostly from single-camera perspectives, while live soccer broadcasts involve dynamic switching between multiple cameras, introducing variations not accounted for in the training phase. In sec. III, we describe our methodical improvements to address this challenge.

III. Methodical and Technological Improvements for Better Event Detection Outcomes

A. Integrating Multi-Class Image Classifier

In order to tackle the challenge of dynamic broadcasting nature of soccer games and improve events detection performance, we introduce multiple modifications to the events detection and classification process. Based on [6], we develop and employ a Multi-Class Image Classifier that distinguishes an input frame into one of the following classes: Close-Up frame; Overview frame; and Transition frame, examples of which are displayed in Figures 2(a), 2(f), and 2(e), respectively. The model was trained for 10 epochs with the resulting average Accuracy of 97.4% on the testing data. We integrated this model into our pipeline to segment the input video into separate streams: those consisting of only the Close-Up shots, and those consisting of only the Overview shots, as well as a set of videos each for a replay shown by the broadcasters during the game.

B. Integrating YOLO

The soccer broadcast videos typically contain additional visual information, such as a scoreboard. To take advantage of this information in order to improve the detection performance, we leverage the YOLOv8 model. Figure 2(b) shows an example of how YOLO detects booking and substitution events portrayed by the broadcaster via additional textual information in the bottom part of the screen. The input video is sent to the YOLOv8 model, where it processes one in 10 frames to look for the events of Substitution, Booking, and to track the scoreboard that is updated every 10 seconds, i.e. once for every 250 frames.

The model is trained for 200 epochs and achieves an average Precision of 91.2% with an average Recall of 92.5% on the test set. YOLO demonstrates the following class-wise Precision for various classes: Booking – 89.7%; Scoreboards – 98.14%; Substitution – 88.9%; and the average for all these classes results in 92.25%. In Table II, we compare the performance of our event detection and classification approach with the approaches developed in other papers. It is important to note that the other approaches mentioned in Table II perform the detection and classification based on the player actions and positions, while our model employs only the additional information displayed by the broadcaster during the game. The comparative analysis demonstrates the superiority of our proposed approach, when the additional information is available, in detecting substitution and booking events over traditional events detection or action classification methods. Another advantage is that our approach requires much less training data.

TABLE II
Precision showed by YOLO in detecting Substitution and Booking events by using additional broadcasting information compared to other approaches

Model	Training Images	Testing Images	Precision
Substitution event			
CALF-60-s [7]	1708	562	60.07
CALF-60-20 [7]	1708	562	46.70
CALF-120-40 [7]	1708	562	67.35
YOLOv8 (our paper)	**180**	**35**	**88.9**
Booking event			
B-CNN [5], [12]	5500	500	66.86
OSME + MAMC using ResNet50 [5], [11]	5500	500	61.70
OSME + MAMC using EfficientNetB0 [5], [12]	5500	500	79.9
YOLOv8 (our paper)	**170**	**34**	**89.7**

Since the employed YOLO model demonstrates high performance in identifying the Booking and Substitution events, and extracts the information from the scoreboard accurately enough, we narrow down the set of key events detected and classified by VideoMAE and delegate these events to be detected by YOLO. The motivation behind this is the reduction of classes needed to be selected for the event by VideoMAE, which appears to be an effective measure for improving the detection and classification performance. The new dataset for VideoMAE consists of the following events: Transition; Set-Piece; Penalty; and Open-Play. The remaining events that initially were in the training dataset (Substitution and Booking) are now processed by the YOLO model. Other removed events include Goals and Celebration, which can be

TABLE III
COMPARISON OF THE OFFICIAL HIGHLIGHT VIDEO WITH THOSE GENERATED BY OUR PIPELINES

		Developed pipeline versions	
Video information	**Original highlight**	**P1**	**P2**
Highlight length	2:07	12:20	10:35
Goal	2	2	2
Foul	2	0	2
Free-kick	0	4	3
Corner	2	7	5
Card	1	5	5
Substitution	1	4	4
Attacks	0	1	18
False-Positives / Duplicates	N/A	12	6

detected by further processing the video sequences close to the scoreboard changes detected by YOLO. The final dataset for VideoMAE comprises around 150 videos for each category. After re-training VideoMAE for 40 epochs, the model reaches impressive Accuracy for the classes it was tested on: Open-Play – 98.7%; Penalty – 91.67%; Set-Piece – 92.86%; Transition – 100%; and demonstrates the average over these classes of 95.34%.

IV. Post-Processing and Highlights Generation

The Multi-Class Image Classification module employs a flag to monitor the current state of the video, distinguishing between live and replay modes. Depending on the classified category of the frame, it is accordingly appended to either the Close-Up or Overview video segments. The detection of a Transition frame triggers a change in the system's state: the flag is set to true, signifying the commencement of a highlight sequence. Subsequent frames are then added to the generated highlight until the system identifies another Transition frame. Upon this detection, the flag reverts to false, indicating the end of the highlight sequence, and the process resumes from the beginning.

After dividing the video into the Close-Up and Overview segments, the Close-Up ones are discarded as they are one of the primary reasons for false classifications since they do not really provide any useful information for the model to detect corresponding events. The extracted Overview segments are processed by the VideoMAE classifier to identify and discern events within the video. Simultaneously with the previous steps, the input video is processed by the YOLOv8 model, which processes each 10th frame to look for the Substitution and Booking events. The scoreboard tracker is updated every 10 seconds.

V. Evaluation on Real Case with Official Highlights

In order to test the performance of the developed pipeline, we compare the highlights generated by our pipeline with the official real highlights released by one of the broadcasters after the game. We selected highlights published for the Premier League[4] game from the Year 2014-15 between Chelsea and Burnley, happened on Feb 21, 2015, and ended with a scoreline of 1-1. To evaluate if the usage of additional information – replays shown by a broadcaster during the game – can be employed to improve the highlights generation performance, we deploy and evaluate two distinct versions of our pipeline. Each version functions as follows:

The first pipeline version (P1) employs a Multi-Class Image Classifier to separate the input video into two categories: Overview and Close-Up, without extracting the replay segments.

The second pipeline version (P2) advances this process by extracting the replay segments in addition to dividing the input video into Overview and Close-Up videos, and adding the extracted replays to the final highlights generated.

The official highlights provided by the broadcaster after the game comprised 2 minutes and 7 seconds of video footage. After processing the game recording, P1 generated the 12 minutes and 20 seconds of highlights, while P2 produced a slightly shorter compilation, totaling 10 minutes and 35 seconds. The information regarding the extracted events can be found in Table III. While original highlights included the critical events affecting the score, it omitted several other significant plays that could provide a fuller narrative of the game. P1 generated highlights that included some repeated events, as it did not extract the broadcaster's replay segments. This resulted in the duplication of two corners and one free-kick within the highlights generated. P2 addressed this by using broadcaster replays as a foundation, augmenting them with events detected by YOLOv8 and VideoMAE to produce comprehensive and informative highlights.

Although both pipelines were able to effectively keep track of the Goal, Substitution and Booking events, there were a few false positives. The first pipeline resulted in 12 false positives, while the second pipeline was able to reduce this number to 6. These inaccuracies occurred during the Open-Play, where the positioning of players and the game's momentum bore resemblance to scenarios typically involving corners and free-kicks. The Multi-Class Image Classifier demonstrated high performance in isolating in-game replays presented during live broadcasts. In our practical test the model successfully identified and extracted all 17 replays showcased during the game, utilizing the transition phase as a key indicator. Besides, the model was also able to effectively separate the input video into various segments: a video consisting of only the Close-Up frames and a video only consisting of Overview frames with the Accuracy of 97.5% using a confidence threshold of 98% that used to reduce false positives. We tested VideoMAE on the Overview video segments extracted by Multi-Class Image Classifier. It demonstrated a commendable Accuracy of 66% on average. Specifically, the model accurately detected 7 Set-Piece events, which are basically challenging to detect and included free-kicks and corners. Table IV refers to the

[4] https://www.premierleague.com/

TABLE IV
SUMMARY OF THE EMPLOYED ML MODELS, AMOUNT OF TRAINING DATA, ML TASKS, AND PERFORMANCE RESULTS DEMONSTRATED BY THE MODELS IN THE DEVELOPED PIPELINE

Task	**Model**	**Training data**		**Type**	**Accuracy/Precision**
		Train	**Test**		
Overview frames	Multi-Class Image Classifier	550 Images	150 Images	Image classification	97%
Close-up frames	Multi-Class Image Classifier	500 Images	130 Images	Image classification	95%
Transition (Logo) Frames	Multi-Class Image Classifier	540 Images	140 Images	Image Classification	99%
Booking detection	YOLOv8	175 Images	75 Images	Object detection	89.7%
Substitution detection	YOLOv8	175 Images	75 Images	Object detection	88.9%
Goals detection	YOLOv8 and AWS Rekognition	550 Images	140 Images	Object Detection and Text Extraction	98.14%
Set-Piece detection	VideoMAE	140 Videos	40 Videos	Video Classification	92.86%
Penalty Detection	VideoMAE	120 Videos	35 Videos	Video Classification	91.67%

comprehensive overview of various tasks within the developed pipeline, detailing the models employed, the amount and the division of training and testing data, the type of ML task for each action, and the corresponding performance achieved.

VI. CONCLUSION

In this paper, we developed and presented a novel approach to automate key event detection and highlights generation from soccer match videostreams. We implemented a software pipeline that leverages extracting additional visual information provided by broadcasters, and integrates intelligent video analytics to detect events such as bookings, substitutions, and goals. This approach outperformed existing methods that rely solely on analyzing the soccer scene, demonstrating superior performance and requiring less data for training. In our study, we presented the following contributions: **(1)** we developed an intelligent soccer events detection and highlights generation pipeline and made it available to the public; **(2)** to enhance events detection and classification performance, we developed and integrated software module for segmenting the videostream into distinct frames corresponding to various camera angles and activities (repeats of events or actual gameplay), processing these segments in separate flows; **(3)** we further improved events detection and classification performance by fusing intelligent video analytics performed by VideoMAE with employing YOLO to process additional visual information added by broadcasters to their streams. This allowed us to generate more comprehensive and succinct highlights. We evaluated our pipeline on a real soccer match recording and compared the highlights it generated with the official ones provided by the broadcaster. Our pipeline demonstrated ample performance in event detection and highlights generation, successfully identifying all key events in the game. Our approach offers significant benefits to broadcasters, sports analysts, and content creators by reducing the manual effort and expertise required for highlights generation. The ability to automatically detect and highlight key moments in a soccer match not only saves time and costs, but also ensures a comprehensive and engaging viewing experience for soccer enthusiasts.

REFERENCES

[1] A. Cioppa, A. Deliege, F. Magera, S. Giancola, O. Barnich, B. Ghanem, and M. Van Droogenbroeck, "Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4537–4546.

[2] N. Darapaneni, P. Kumar, N. Malhotra, V. Sundaramurthy, A. Thakur, S. Chauhan, K. C. Thangeda, and A. R. Paduri, "Detecting key soccer match events to create highlights using computer vision," *arXiv preprint arXiv:2204.02573*, 2022.

[3] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1711–1721.

[4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[5] A. Karimi, R. Toosi, and M. A. Akhaee, "Soccer event detection using deep learning," *arXiv preprint arXiv:2102.04331*, 2021.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[7] O. A. Nergård Rongved, M. Stige, S. A. Hicks, V. L. Thambawita, C. Midoglu, E. Zouganeli, D. Johansen, M. A. Riegler, and P. Halvorsen, "Automated event detection and classification in soccer: The potential of using multiple modalities," *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 1030–1054, 2021.

[8] A. Raventos, R. Quijada, L. Torres, and F. Tarrés, "Automatic summarization of soccer highlights using audio-visual descriptors," *SpringerPlus*, vol. 4, no. 1, p. 301, 2015.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[10] L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly, "End-to-end camera calibration for broadcast videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 627–13 636.

[11] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 805–821.

[12] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.