SoccerNet 2024 Multi-View Foul Recognition WJB Team Technical Report

Jing Zhang, Xinyu Liu, Kexin Zhang, Yuting Yang, Licheng Jiao, Shuyuan Yang
Intelligent Perception and Image Understanding Lab, Xidian University
jingzhang_work @163.com

Abstract

The SoccerNet 2024 multi-view foul recognition challenge requires classifying the severity of the foul and the action of the foul based on football foul videos from multiple perspectives to help referees make decisions. In this work, we propose a multi-view video fusion classification network——M2VFCN. We use Vit-B as Backbone, use the pre-weights of the unmasked teacher foundation model [6] for initialization, and use max pooling for feature fusion. In order to achieve a trade-off between training time and training effect, we extract 8 frames of images from each video for training. Finally, our method achieved a challenge stage score of 44.76 in the SoccerNet 2024 Multi-View Foul Recognition Challenge.

1. Introduction

The SoccerNet 2024 Multi-View Foul Recognition challenge aims to help referees make decisions. The use of automated decision-making systems can better reduce the referee's burden, save manpower, material and financial resources, and improve the accuracy of referees. [4] proposed SoccerNet-MVFouls, a new multiview video dataset. And proposed the VARS system, a multi-camera video recognition system for classifying fouls and their severity.
In this work, we used SoccerNet-MVFouls and proposed a multi-view video fusion classification network (M2VFCN), aiming to better help referees make more fair and equitable decisions through multi-view videos.

2. Related Work

Sports video understanding. This field has grown due to its challenging nature, leveraging deep learning for tasks like player detection, tracking, tactics analysis, and player re-identification. Initially centered on video classification for action recognition and game phase differentiation, the focus has shifted to temporal activity localization and action spotting. Progress relies on large-scale datasets like SoccerNet, which provides extensive benchmarks and fosters research through annual competitions.
Video understanding. Historically hindered by a lack of large datasets, video understanding has advanced with datasets like UCF101 [8], ActivityNet [3], YouTube-8M [1], and Kinetics [5]. Key tasks include video classification, action recognition, captioning, and generation. Techniques like Temporal Segment Network (TSN) and spatio-temporal convolutional blocks (R(2+1)D) have improved performance. Recently, Multiscale Vision Transformers (MViT) [2, 7] have combined CNNs and transformers to capture spatial and temporal attentions. This work trains various video representations to learn per-clip features, aggregating them from multiple views to identify foul properties in soccer.

3. Methodology

In this work, we propose a Multi-View Video Fusion Classification Network (M2VFCN). As shown in Figure 1, we use ViT-base as the backbone and initialize it with pre-trained weights from the unmasked teacher model [6]. Feature fusion is performed using max pooling. Noting the class imbalance among different categories, we employ a weighted loss for training. To balance training time and effectiveness, we extract 8 frames from each video for training.

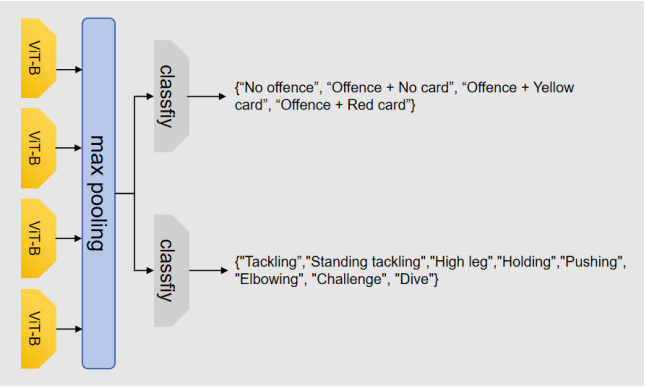

Figure 1. Overview of M2VFCN architecture.

4. Experiment

4.1. Experimental setup

Training details. For both classification tasks, we leverage clips of 8 frames, with a spatial dimension of 256×256 pixels. Specifically, the clips contain 4 frames before the foul and 4 frames after the foul.The encoders ViT-B are pre-trained as detailed in [4], and the classifier C is trained from scratch, while both are trained in an end-to-end fashion. We use a cross-entropy loss, optimized with Adam with an exponential decreasing learning rate starting at 2e-5 and the total batch size is 24. We achieved our current score using the 29th epoch, and it takes around 1.5 hours to train on a 6 Nvidia V100 GPU.

Evaluation metrics. We report classification accuracy, defined as the ratio of correctly classified actions to the total number of actions. We also provide top-2 accuracy, where a sample is considered correctly classified if its action appears in the top two highest confidence predictions, to gain deeper insights into the model's performance. Due to the imbalance in the dataset, we also report balanced accuracy, defined as follows:

$$BalancedAccuracy(BA) = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{P_i} \qquad (1)$$

with N the number of classes, TP (True Positives) is the number of times where the model correctly predicted the class i and P_i (Positives) is the number of ground-truth samples for that class in the dataset.

4.2. Main Results

Task 1: Fine-grained foul classification.
Using our M2VFCN network, we attained an Accuracy Action metric of 48.86877828054298 and a Balanced Accuracy Action metric of 30.877219531880552 in Fine-grained foul classification. We also employed Uniformerv2 and mvit_v1_b as the backbone of the networks, but their performance was not as good as our method.

Task 2: Offence severity classification.
Using our M2VFCN network, we attained an Accuracy Offence Severity metric of 48.86877828054298 and a Balanced Accuracy Offence Severity metric of 30.877219531880552 in Fine-grained foul classification.

5. Conclusion

This work presented M2VFCN, a model designed to address the task of multi-view foul recognition in soccer matches. In our experiments, we addressed the issue of imbalanced data and took into account the efficiency of training. As a result, we achieved promising results with fewer training resources.Our method achieved a challenge stage score of 44.76 in the SoccerNet 2024 Multi-View Foul Recognition Challenge.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016.

[2] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[4] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. Vars: Video assistant referee system for automated soccer decision making from multiple views, 2023.

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[6] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19891–19903, 2023.

[7] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection, 2022.

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012.