**Football Event Detection and Player Tracking System**

**Muhammad Talha 21K-3288**

**Ibad Ahmed 21K-4660**

**Objective**

This project aims to develop a comprehensive computer vision system that automatically detects and classifies football events while tracking players and the ball in video footage. By integrating YOLOv5 for object detection with temporal event classification models, the system can identify 16 distinct football events including Ball out of play, Clearance, Corner, Direct free-kick, Foul, Goal, Indirect free-kick, Kick-off, Offside, Penalty, Red card, Shots off target, Shots on target, Substitution, Throw-in, and Yellow card with high temporal precision.

**Problem Statement**

The manual analysis of football matches presents numerous challenges that limit the effectiveness and scalability of traditional approaches. Professional analysts typically spend hours reviewing match footage to identify key events, track player movements, and generate statistics. This process is not only time-consuming and labor-intensive but also prone to human error and subjective interpretation. As the volume of matches increases, manual analysis becomes increasingly impractical, creating a bottleneck in the sports analytics pipeline. Furthermore, the demand for real-time insights during live broadcasts and coaching sessions requires processing speeds that human analysts cannot achieve.

The technical challenges of automating football video analysis include accurately detecting multiple players and the ball across varying camera angles, identifying precise moments when specific events occur, and distinguishing between visually similar events while maintaining computational efficiency for practical deployment.

**Methodology**

Data Acquisition and Preprocessing

We utilized the SoccerNet dataset, which contains hundreds of broadcast soccer matches with precise temporal annotations for key events. The raw video footage was processed to extract individual frames at consistent intervals, creating a frame-based representation suitable for deep learning models. To enhance model generalization and robustness, we implemented extensive data augmentation techniques including random crops, color jittering, and brightness adjustments. We also employed mixup augmentation and implemented foreground upsampling to address the significant class imbalance inherent in football event detection.

**Object Detection with YOLOv5**

Player and ball tracking form the spatial foundation of our event detection system. We implemented YOLOv5, a state-of-the-art object detection framework, to identify and localize players and the ball in each frame. YOLOv5 was selected for its excellent balance of accuracy and computational efficiency, capable of processing frames in real-time while maintaining high detection precision. The model was fine-tuned specifically for football scenarios, adapting to the unique challenges of tracking fast-moving objects in wide-angle broadcast footage. We configured custom anchor boxes, implemented confidence thresholding, and applied non-maximum suppression algorithms to enhance detection performance in football contexts.

**Feature Extraction Architecture**

Our feature extraction pipeline employs a flexible backbone architecture that can be configured with different CNN models based on deployment requirements. For scenarios prioritizing accuracy, we implemented ResNet-50, while RegNetY models offered a favorable accuracy-to-computation ratio for more efficient processing. These backbone networks were enhanced with temporal modeling capabilities through the integration of specialized modules like the Temporal Shift Module (TSM) and Gate Shift Module (GSM), which enable information exchange between adjacent frames. The feature integration component combines detection-based features from YOLOv5 with appearance and motion features from the backbone networks, creating a comprehensive representation that encodes both spatial configuration and movement patterns.

**Temporal Modeling**

Our temporal modeling approach employs Gated Recurrent Units (GRUs) to process sequences of frame features and identify temporal patterns indicative of specific events. For standard deployments, a single-layer GRU provides excellent performance, while more complex scenarios benefit from a deeper three-layer configuration. We also explored alternative temporal architectures including Multi-Stage Temporal Convolutional Network (MS-TCN) and ASFormer, which leverage different approaches to capture temporal dependencies. The event classification component processes these temporal features to identify specific football events among the 16 target classes, with class weighting implemented to handle the significant class imbalance inherent in football event detection.

**Training Pipeline**

Our training methodology employs mixed precision training using PyTorch's Automatic Mixed Precision (AMP), gradient accumulation techniques to expand effective batch sizes, and a learning rate scheduling approach that combines linear warmup with cosine annealing. The optimization process utilizes AdamW with weight decay regularization, and for multi-GPU environments, we implemented DataParallel training to distribute the workload across available hardware. Our checkpoint management system regularly saved model states throughout training, enabling easy resumption after interruptions and preservation of the best-performing model configurations.

**Results**

SoccerNet Action Spotting Performance

• Test Split Performance:

- Average-mAP (1-5s tolerance): 61.19% with 200MF model, 61.82% with 800MF model

- Average-mAP (5-60s tolerance): 73.25% with 200MF model, 74.05% with 800MF model

• Challenge Split Performance:

- Visible Events (Shown):

- 70.41% average-mAP with 200MF model (1-5s tolerance)

- 72.76% average-mAP with 800MF model (1-5s tolerance)

- 14.1 percentage point improvement over previous state-of-the-art

- Non-Visible Events (Unshown):

- 45.98% average-mAP with 200MF model (1-5s tolerance)

- 51.65% average-mAP with 800MF model (1-5s tolerance)

- 2022 SoccerNet Challenge:

- 66.73% overall average-mAP (2nd place)

- Only 1.1 percentage points behind the winning submission

- Superior performance on visible events with 74.84% average-mAP

**References:**

1. Hong, J., Zhang, H., Gharbi, M., Fisher, M., & Fatahalian, K. (2023). Spotting Temporally Precise, Fine-Grained Events in Video. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

2. YOLOv5: Ultralytics. (2020). YOLOv5. GitHub repository. https://github.com/ultralytics/yolov5

3. SoccerNet: Giancola, S., Amine, M., Droit, T., & Ghanem, B. (2018). SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. arXiv preprint arXiv:1804.04527.

4. Deliege, A., Cioppa, A., Giancola, S., et al. (2021). SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. CVPR Workshops.

5. Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal Shift Module for Efficient Video Understanding. ICCV 2019.