Football Analysis System & Team Identifier using YOLO11 and CLIP

LIS4042 Artificial Vision — Universidad de las Américas Puebla

Problem Overview

Modern football, faces a critical challenge in the gap between broadcast audiences and live audiences. While broadcast audience benefit from advanced computer vision analytics and augmented visualizations, in-stadium fans experience the match with limited real-time insights. This creates an engagement gap:

Information gap: Surveys show that 58% of fans wish they had access to the same statistics, analysis, and replays at the stadium as at home

Reduced engagement: Fans attending live games miss advanced insights (possession, sprint speeds, shot maps) that TV audiences enjoy.

The project aims to show reliable detection and identification, and identification can be done using only a single neural networks and CV techniques.

Solution Architecture

A football analysis system was developed using DL and ViT to be able to calculate ball tracking, team identity, tactical positioning and. The pipeline integrates:

- YOLOv11 : State of the Art multiclass object detection for real-time scenarios
- CLIP ViT : Zero-shot ViT chosen for making strong semantic relationships with images
- K-means — Unsupervised Machine Learning model to classify CLIP embeddings
- UMAP— For dimensionality reduction from CLIP

The system excels at Player identification as it leverages cv techniques such as : Sharpening filter, Denoising and contras enhancing. Team classification as it uses CLIP that creates 768-dimensional embeddings, which is then reduced using UMAP, making a state-of-the-art clustering pipeline

Neural Network Detection (YOLOv8)

YOLO11 detects 4 classes: Players, goalkeeper, refeeres and balls. The model was trained in 50 epochs with batch_size = 6 and 960x960 images. The dataset had over 1000 images; the model performance was the following:

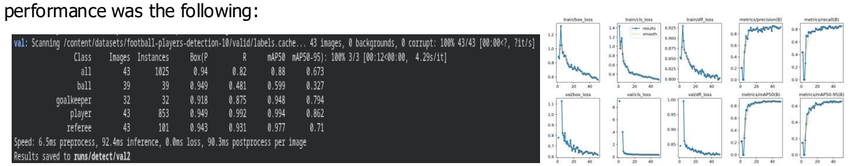

Figure 1. YOLO11 model results.

As observed the mAP of the model is 0.88. Having some issues in the ball detection. This is due to the small size of the ball in a reduced image, that makes it hard for the model to always detect the ball, as it occupy a small number of pixels. Nevertheless, the model outperforms other models using the same dataset with previous YOLO versions by a high margin as seen in image 2 . CLIP and k-means will help to boost True positive detections by marking the difference between goalkeepers, players and referees,

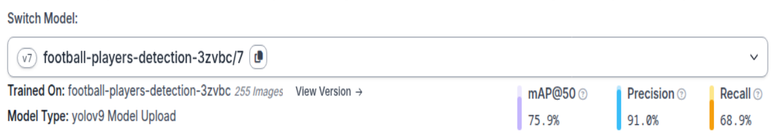

Figure 2. Yolo9 performance metrics with same dataset .

Key Contributions

- Full single-camera system estimating player and ball speed.
- Combination of YOLOv11 + CLIP + Classical CV implemented from scratch.
- Unsupervised team clustering + UMAP + K-means.
- Open-Source system to generate tactical insights with external data
- Ablation study comparing

Vision Transformer Team Classification

CLIP is a model developed by openAI that produces 768-dimensional embeddings for each detected player crop. CLIP has an "image encoder and text encoder to get visual features and text features" (Huggingface, n.d.)

Pipeline:

1. YOLO crops (960×960)
2. CLIP embedding (1,768)
3. UMAP dimensionality reduction (1,3)
4. K-means (k = 2) -> Because we need to classify in a binary was

For Classifying the K-means model we used the following metrics:

Silhouette: 0.81 -> Each player crop is correctly matched with each cluster

Davies-Bouldin : 0.27 -> Each cluster is separated In a very good way

Calinski-Harabasz : 5286.48 -> Good ratio of between-cluster variance to within-cluster variance. Showing very dense clusters

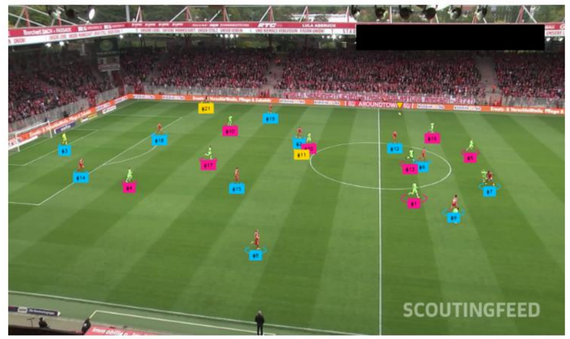

Figure 3. Yolo11 detection with CLIP and kmeans team identification.

Ablation Study

There are other methods that are used for team identification like extracting the mean value of the RGB of each crop. Although this method is faster that CLIP, it does not perform as good. We can see the graphs to analyze the clustering performance of each method

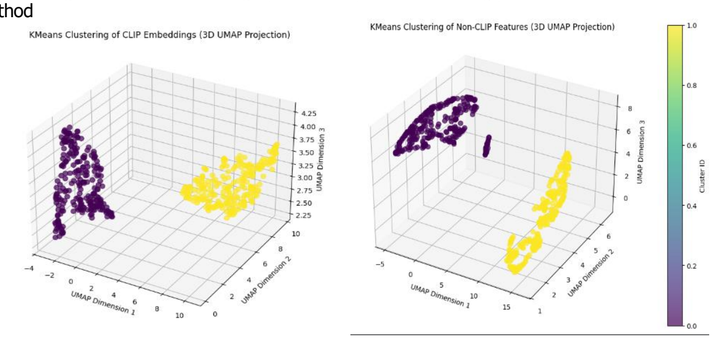

Figure 4.. On the left: clustering with CLIP , On the right, clustering with RGB

As we can see in iimage 4, we get considerably better results with the CLIP embeddings, having the following results with pur RGB :

METHOD	Silhouette	Davies-Bouldin	Calinski-Harabasz
CLIP	0.81	0.27	5286.4
RGB	0.83	0.69	829.53

Final System Output

The system produces the detection:

By detection player boxes with YOLO11

- Uses RGB extraction for fast identification
- Create CLIP embeddings every 3 seconds to
- Feedback the RGB extraction
- Uses Kmeans for identification
- Calculates centroid distance for better
- Goalkeeper identification

Figure 4. Final analysis visualization.

Discussion

Strengths

- Accurate player detection and team clustering.
- Real-world identification from a single broadcast camera with RGB and CLIP clustering.
- Modular pipeline usable in real-time scenarios.

Limitations

- Ball detection is low due to image size limitations.
- Identification becomes hard fast pans or continous camera movements.
- CLIP is quite slow as it needs to generate embeddings each time new frames are passed

Future Work

- ML-based field keypoint detector for more on-device insights about the game
- Better ball detector for small-object accuracy. Camera stabilization methods

References

1 *CLIP*. (s. f.). https://huggingface.co/docs/transformers/model_doc/clip

2 OpenCV. (2025, November 22). *OpenCV - Open Computer Vision Library*. https://opencv.org/

3 Ultralytics. (2025, November 23). *Home*. https://docs.ultralytics.com/es/