The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 6, December 2025
https://doi.org/10.33022/ijcs.v14i6.5039

Hybrid-Based Multi-Object Tracking for Football Sport

Zin Mar Htun[1], Theingi Myint[2]
zinmarhtun.myanmar@gmail.com[1], drtgim@gmail.com[2]
[1] Computer Engineering and Information Technology, Technological University(Taunggyi), Myanmar
[2]Computer Engineering and Information Technology, Mandalay Technological University, Myanmar

Article Information

Received : 22 Nov 2025
Revised : 29 Nov 2025
Accepted : 3 Dec 2025

Keywords

MOD, YOLOv8n, FairMOT, stacked LSTM, MOTA, MOTP,mAP.

Abstract

Tracking is now popular in real world. Precise tracking of objects in real-time videos is a challenging task. With billions of fans, football is a rapidly expanding sport that has proven essential to many nations and their citizens in particular. None of the numerous great target tracking algorithms have surfaced in recent years primarily deep learning and correlation filtering that can track players in soccer game videos with high accuracy. In this paper, the proposed system is used You Only Look Once version 8-nano (YOLOv8n) for Multi-Object Detection (MOD) to get higher detection accuracy results. Moreover, this system is based on the hybrid method for tracking. The hybrid method is combined with stacked Long Short Term Memory (LSTM) and Fairness of Detection and Re-identification in Multipe Object Tracking (FairMOT). The experimental analysis shows that the proposed system is efficiently and better accuacy because the best detection results with YOLOv8n is 93% for precision, 91% for recall and 92% for mAP(50) with own dataset. After using the proposed system, the average of the Multi Object Tracking Accuracy (MOTA) is 80 % at IoU-Threshold 0.5, the average of the Multi Object Tracking Precision (MOTP) is 89% at IoU-Threshold 0.8 and the average of the final mAP is 96% at IoU-Threshold 0.5 by using hybrid method for tracking.

A. Introduction

High-level semantic activities are included video summary creation, player motion analysis, game strategy formulation, and football event recognition. All of these activities are based on target tracking technology, which is extremely important in football game videos. It is examined a deep learning-based soccer player monitoring method. To improve the algorithm's ability that can identify ball, player, referee, site referee, staff and goal keeper, a deep learning based football player tracking scheme is proposed. A convolutional neural network is constructed to extract the rich visual features of players in football game videos, and the network is trained on numerous data sets containing similar objects [1].

Jiating Jin proposed multi-object tracking is a middle-level task in computer vision, which plays a crucial role in many fields, for example, unmanned driving, video surveillance, and the robotics control in 2021. Its main purpose is to achieve the portrayal of the object trajectory in the video sequence. There are some challenges in the process of tracking, such as serve occlusion, deformation of objects, complex environment, similar appearance etc. In order to solve the mentioned problems, some methods have been proposed, one of which is the hierarchical data association [2].

Shuo Wang proposed and explored object detection frameworks based on deep learning in 2021. The foundation of conventional object detection technique shallow trainable structures and handmade features. With the rapid advancement of deep learning, intricate models that combine both low-level image attributes and high-level context from object detectors and scenes, capable of learning semantic and deeper features, are being introduced [3].

The main objectives of the proposed system is to collect and preprocess a football sport video dataset, to create the own dataset from football sport video, to implement YOLOv8n for real time object detection, to design and train a stacked LSTM network with FairMOT that can predict the future prositions of players and the ball based on their historical trajectories in football matches and to develop a tracking algorithm that can associate and maintain the identity of players and the ball across frames using the object detection results and stacked LSTM predictions.

The system is proposed a multi-object tracking algorithm with stacked Long Short-Term Memory, which uses the stacked LSTM structure to establish an encoder-decoder network as the motion model of objects. The historical information of trajectories is applied to predict the possible positions in the future of objects. FairMOT is based on the anchor-free object detection architecture CenterNet and performs the state-of-the-art methods by a large margin on several public datasets. FairMOT tracker can perform object detection and re-identification and the result achieves high accuracy for both detection and tracking. The stacked LSTM network can have a better performance in tracking through the accurate predictions of locations. Meanwhile, in order to reduce missed targets and the tracking trajectories are clustered and optimized, the accuracy of tracking can improve.

The main contributions of this paper are as follow:

1. FairMOT tracks player based on current frame only, might lose track using occlusion and then it can not predict where the ball is going. To solve this problem, the proposed system is used by integrating with stacked LSTM.

2. Stacked LSTM predicts player movement even during occlusion, ball trajectory forecasting improves from false tracking, smooths and interpolates trajectories over time using learned patterns, encode complex behaviors (eg: player positioning strategies and movement).
3. There is currently no existing study that employs a Stacked LSTM in conjunction with FairMOT for Multi Object Tracking.
4. End-to-End Pipeline Integration, where detection, preprocessing, sequence modeling, and prediction are unified into a single framework.
5. Experiments are conducted on the own dataset, which proves the proposed algorithm can achieve a better performance in Multi Object Tracking Accuracy (MOTA) and Multi Object Tracking Precision (MOTP).
6. Application to complex multi-object scenarios (e.g., sports video analysis, multi-agent tracking), demonstrating adaptability beyond standard benchmarks.

The aim of the proposed system is to accurately detect and track all relevant multiple objects in each frame in a video sequence, including players, goalkeepers, football, main referee, side referees, and staff.

The structure of the paper is as follows. Section B describes the overall structure of the presented method in detail, including stacked LSTM, the two-stage tracking framework the network flow data association and FairMOT tracker. Section C demonstrates and discusses the experimental results on the own dataset. Section D describes the conclusion of the proposed system. Section E shows the acknowledgement for this paper.

B. Research Method

This section will explain in detail the methodology used in the proposed system and flowchart following this system. The architecture of the method used to detect classes and tracked with hybrid method for football sport is mentioned.

For this system to be complete, firstly, it needs to install dependencies and software following for this system. As this system is used window 10 as platform, Ubantu is installed to the computer. The main programming language for this system is used python. In python, its code is combined YOLOv8n, Stacked Long-Short-Term Memory and FairMOT, to perform the system as multiple classes detection and tracking.

After all dependencies and software are installed in the machine, this system needs to collect images related to football sport to use for training a model for sports detection. The number of classes in the images, the variety of class types in the images, the occlusion images or not occlusion images will affect the accuracy of the model. It will be used because how model learn from the images.

When several different models are ready, the model will run in ubantu by calling python file. From this step, it required user input to the system. Firstly, user can install the dependencies according to what environment need and interest computation for running this system by the user. User can install dependencies in Pip python environment or Conda environment. In this paper, it is used python environment because python can easily creates, saves, loads and switch between environment on the local computer and had installed most basic library. Then we need to setup between CPU or GPU environment. GPU environment should be faster and recommended if computer had one GPU or more. This system is used GPU environment in which will install dependencies for instance Tensorflow-gpu, OpenCV, lxml, tqdm, absl-py, matplotlib, easydict and pillow. The only difference between CPU and GPU environment is TensorFlow library package while other dependencies is same. Next, user need to change the model get from training into TensorFlow model. As the original model YOLOv8n is in darknet framework which written in C, converting darknet model to TensorFlow helps running this system on python in ubantu platform efficiently. Lastly, user can run the object tracker by input specified the type of input video, type of weight use, type of framework, tiny or normal weight type and then trained with stacked LSTM. The training, inference and evaluation will be run in the python which it will describe in the next section.

And then, one of the main part for this system is to get best detection results. Therefore, the proposed system is used YOLOv8n for detection. The "nano" version of YOLOv8, which is a variant optimized for lower computational requirements and faster inference times. The "nano" model is making it suitable for mobile. YOLOv8 is engineered for effective and efficient of object detection models, and fully utilizing modern hardware capabilities. YOLOv8 is anchor-free, it reduces the number of box predictions, accelerating the non-maximum suppression (NMS). YOLOv8 comes in five scaled versions: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra-large). In the field of computer vision, YOLOv8 is transforming real-time object tracking and analysis. It provides a comprehensive overview of the inference pipeline for object tracking and counting using YOLOv8 [4]. We can evaluate the system's effectiveness using several key metrics. Precision and recall are used to measure the accuracy of detecting important match events, ensuring that critical moments (e.g., goals, fouls) were identified with minimal false detections. Additionally, the system's ability to track objects is measured using Multiple Object Tracking Accuracy (MOTA) to assess how well it follows players and the ball over time. The architecture of YOLOv8 is shown in Figure 1.

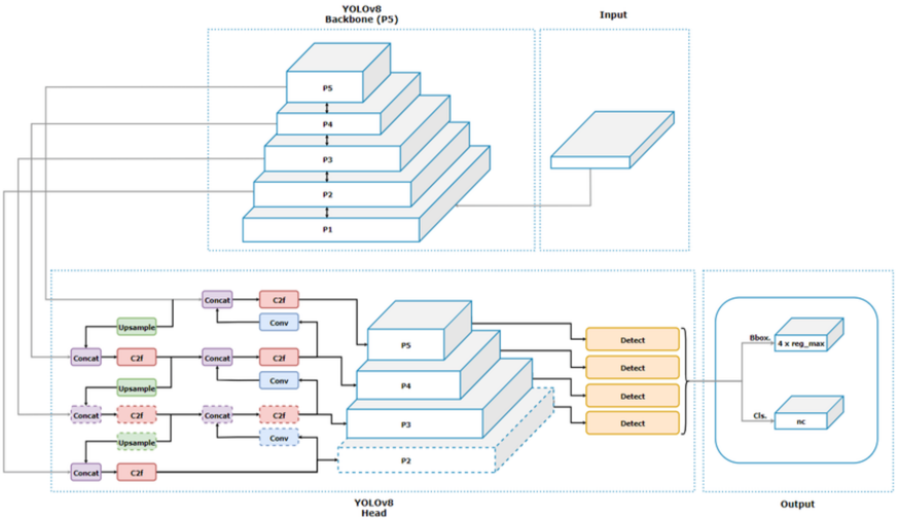

Figure 1. The Architecture of YOLOv8

FairMOT is a Multi Object Tracking (MOT) method that addresses the common "unfairness" issue where object detection is prioritized over re-identification (Re-ID). FairMOT is a one-shot tracker that uses a single network to perform both tasks simultaneously and equally, which increases its accuracy and efficiency. FairMOT is a real-time, anchor-free tracking system that solves identity switch issues by combining object detection and re-identification in a single network, ideal for crowded scenes, surveillance, sports, and autonomous systems [5]. In conventional MOT methods, object detection and re-identification are treated as a two-step process. First, a detector finds objects in each frame. Next, a separate re-ID model extracts features from each detected object to maintain its identity. The architecture of the fairMOT is described in Figure 2.

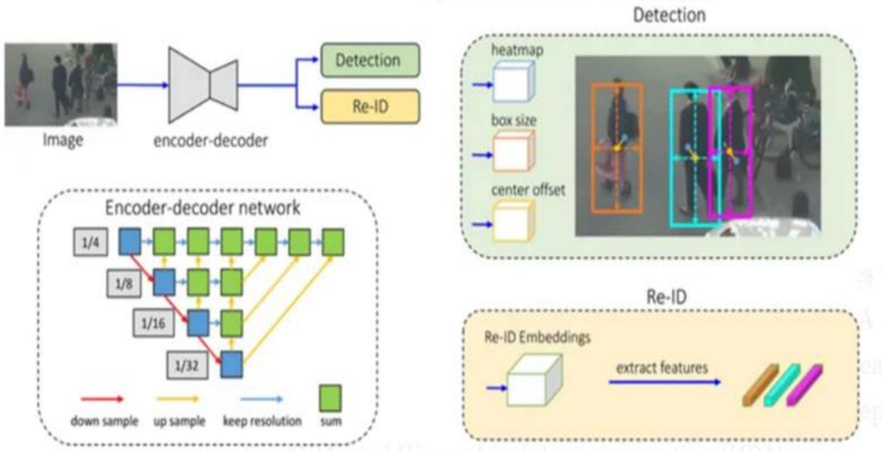

Figure 2. The Architecture of FairMOT

A "stacked LSTM" is an LSTM model with multiple layers of LSTM cells, whereas a standard LSTM has only a single layer of cells. This architecture makes the network "deeper," allowing it to learn more complex, hierarchical representations of sequential data. To create a stacked LSTM, the system needs to configure each LSTM layer to output its full sequence of hidden states, which are used as the input for the next layer. This is typically done by setting a return_sequences parameter to True in machine learning frameworks like Keras or PyTorch. Layer 1 takes the input data sequence and processes it, producing an output sequence of hidden states. Layer 2 takes the output sequence from Layer 1 as its input and performs further processing, creating a more abstract representation. At subsequent Layers, this process repeats for each additional LSTM layer, with each successive layer learning a higher-level, more complex representation of the data. The final LSTM layer typically returns only its last hidden state to a dense output layer for final prediction or classification [6]. The two layers of stacked LSTM is mentioned in Figure 3.

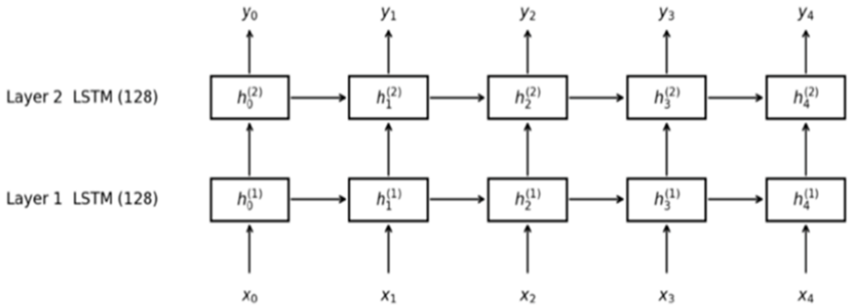
Layer 2 LSTM (128)
Layer 1 LSTM (128)
y_0 y_1 y_2 y_3 y_4
$h_0^{(2)}$ $h_1^{(2)}$ $h_2^{(2)}$ $h_3^{(2)}$ $h_4^{(2)}$
$h_0^{(1)}$ $h_1^{(1)}$ $h_2^{(1)}$ $h_3^{(1)}$ $h_4^{(1)}$
x_0 x_1 x_2 x_3 x_4

Figure 3. Two Layers of Stacked LSTM

Figure 4 shows the methodology flowchart following this system works. Three main modules are dataset creation, detection and tracking module in the proposed system. To collect images of football sport manually by downloading one of football sport video from YouTube videos [7]. After that, images are separated in Google Colab. The proposed system is used 6615 images. Then, the system followed by instruction and understanded the flow of the coding, the image separting from youtube will be run in collab for several hours.

The images are labelled and classified, then the system will be used to train as model in YOLO. Using Google Colab is easier and faster to train model. Google Colab is free, and users enable to write using python language through the browser. It is extremely fast in Colab because it can train model using expensive GPU given by Colab. But too much objects images are not safed to train in Google Colab. Because it had been too long time, and if connection problem is something wrong, it will begin at the start. Therefore, this system is trained on Ubantu.

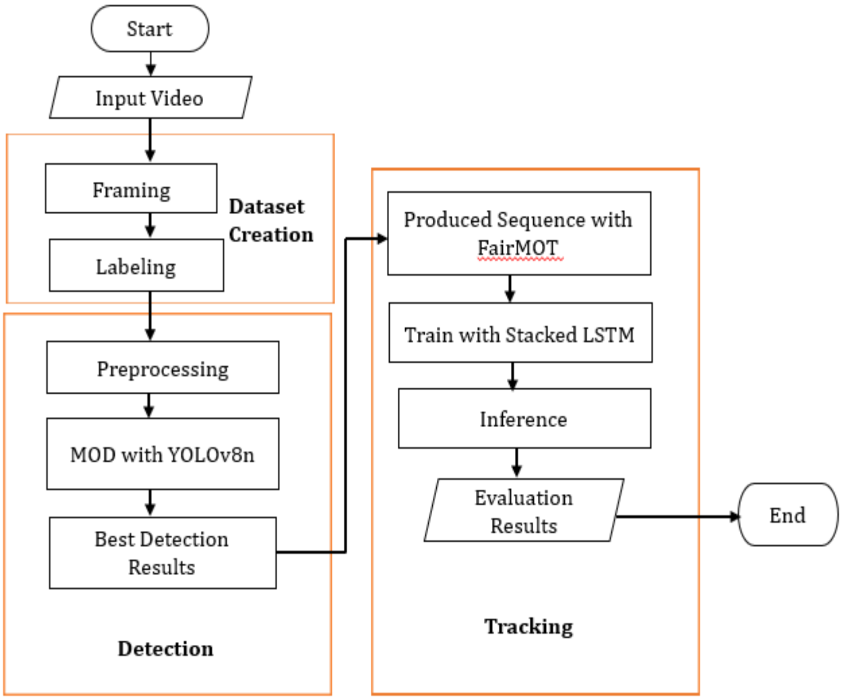

Figure 4. Flow Chart of Proposed System

After collecting a lot of images relating to the football sports, the images need to be label. The labelling and classifying of the images are hassle if it is normally following to YOLOv8n format [8]. In YOLOv8n format .txt-file for each .jpg-image-file must be in the same directory and with the same name. In the txt file, the object number and object coordinate of the related image, the system must be in line like as: <object-class> <x> <y> <width> <height>. Images are labelling and classifying by using Label Studio to do manually and then it is stored in YOLOv8n format in a folder. The implementation of multi-object detection (MOD) with YOLOv8n is illustrated in Figure 5. This proposed system uses six classes in football sport. These classes are identified with IDs by using YOLOv8n for detection process. These IDs are described ball (0), goalkeeper (1), player (2),main referee (3), side referee(4) and staff members (5). The evaluation results of YOLOv8n with own dataset is shown in Figure 6.

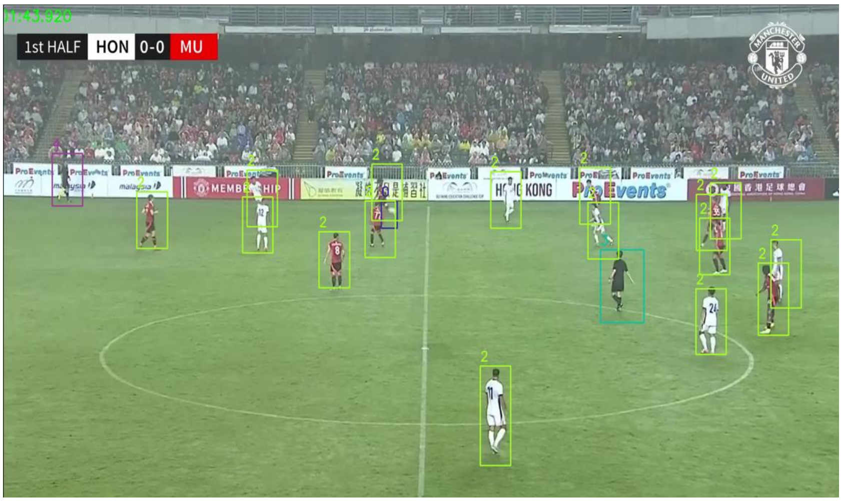

Figure 5. Implementation of Multi-Object Detection with YOLOv8n

Figure 6 shows the evaluation result of YOLOv8n with own dataset. After traing 50 epoches, the performance of detection results are precision (93%), recall (91%) , mAP50 (92%), mAP(50-95) (44%).

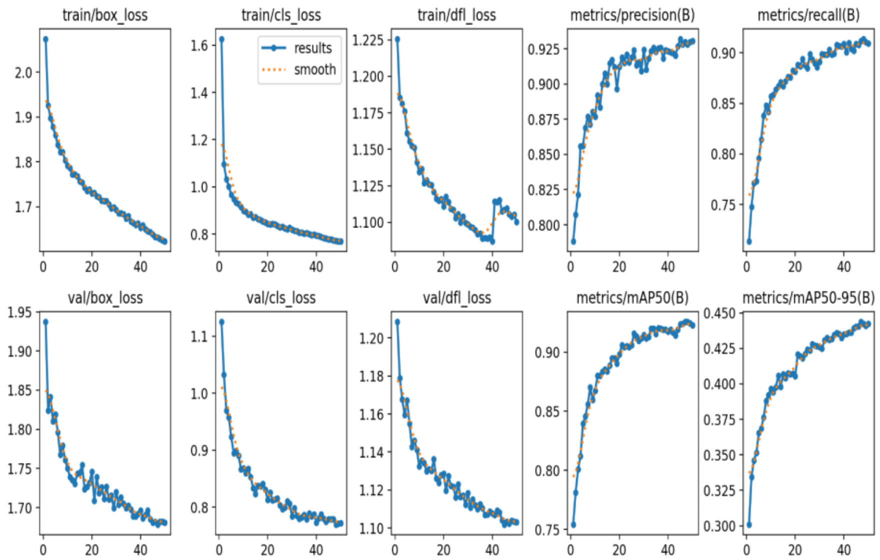

Figure 6. Evaluation Results of YOLOv8n

C. Result and Discussion

Manual inspection are used for YouTube video sequences to obtain the ground truth. The metrics of evaluation are shown in Equations (1), (2), (3), (4) and (5). The proposed system is used 6652 frames for multi-object detection. 5321 frames are used for model train, 666 frames are used for testing and then 666 frames are used for validation. The system is calculated values for true positives, false positive and false negative which can be defined as follows: (i) true positive (TP): ball detected by the model when ball exists in that position; (ii) false positive (FP): ball detected by the model when ball does not exist in that position; (iii) false negative (FN): ball not detected by the model when ball exists in that position.

$$\text{Precision} = TP/\ (TP + FP) \times 100\% \qquad (1)$$

$$\text{Recall} = TP/\ (TP + FN) \times 100\% \qquad (2)$$

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{APi} \qquad (3)$$

Where, N = Number of classes and AP_i = Average Precision for class i.

$$\text{MOTP} = \frac{1}{|TP|}\sum_{TP} S \qquad (4)$$

Where, MOTP is multi object tracking precision, S is similarity Score.

$$\text{MOTA} = 1 - \frac{|FN|+|FP\}+|IDSW|}{|gtDet|} \qquad (5)$$

Where, MOTA is multi object tracking accuracy, IDSW is identity switch and gtDet is groundtruth detection [9].

The system is used with ten videos for tracking, which are downloaded from Kaggle ownership by Shreya Mainkar [10]. This paper describes the total ground truth objects of ten videos respectively. The proposed system is used some parameters for model configuration such as the input size is 6, the hidden size is 300, the output size is 4, the sequence length is 16, the number of LSTM layer is 2, the dropout is 0.3, the width (1920), and the height (1080).

All of these results are obtained after training the new model using hybrid method. The highest multi object tracking precision (MOTP) is 90% for video4 which IoU-Threshold is 0.8 at the proposed system. The highest multi object tracking accuracy (MOTA) is 96% for video5 which IoU_Threshold is 0.5.

At [11], the detection accuracy rate with mAP is 87.98% where the combination of YOLOv4 with the Deep Sort algorithm can detect and track and calculate over 13 types of vehicles using 25000 pictures of training dataset. At [12], custom image dataset is trained six specific classes using YOLO and this model is used in five videos for tracking with SORT algorithm where the best results are 85.20% for accuracy, 95.60% for precision and 93.20% for recall. At [2], LSTM tracking by using Public MOT16 dataset , the accuracy is 60.5%. At [13], the using YOLOv3 and YOLOv4 with deep sort Custom Dataset (7319 images with 4 specific classes: Car, Truck, Bus, Motorcycle and custom video), then the highest accuracy is 82.02% with YOLOv4 while the accuracy is 80% with YOLOv3 (highest accuracy with acceptable speed. At [14], tracking by detection (TBD) by using with 6 datasets (MOT), the results are SORT (34%), BLSTM_MTP (41.3%), respectively. At [15], YOLOv3 (Detection) with SORT (Tracking) by using SoccerNet ISSIA Dataset, which tracking accuracy is 93.7%. At [16], the accuracy is improved from 7.1% to 9.5% (better than YOLOv7) using with tiny person dataset where the detection speed reaches 208 fps.

The proposed system of tracking results are the average of the Multi Object Tracking Accuracy (MOTA) is 80% at IoU-Threshold 0.5, the average of the Multi Object Tracking Precision (MOTP) is 89% at IoU-Threshold 0.8 and the average of the final mAP is 96% at IoU-Threshold 0.5 by using the hybrid method for tracking.

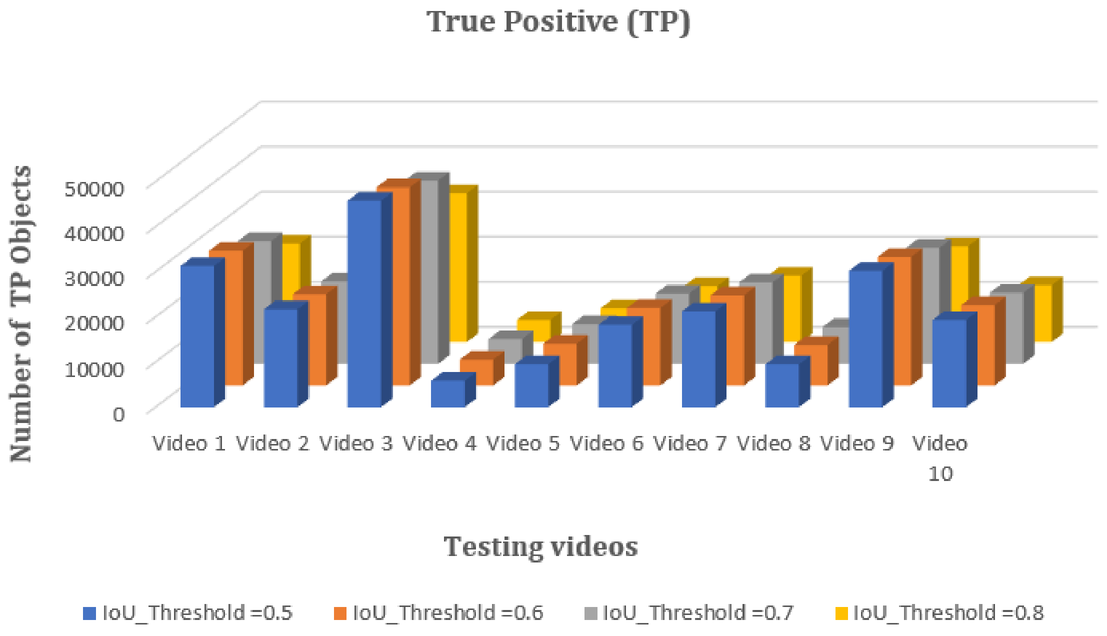

Figure 7. True Positive Values of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 7 shows the number of True Positive (TP) objects of ten videos depending on IoU Threshold (0.5, 0.6, 0.7, 0.8) respectively. The highest of True Positive (TP), which is IoU-Thresholds are 0.5, 0.6, 0.7 and 0.8 at the proposed system, is video 3.

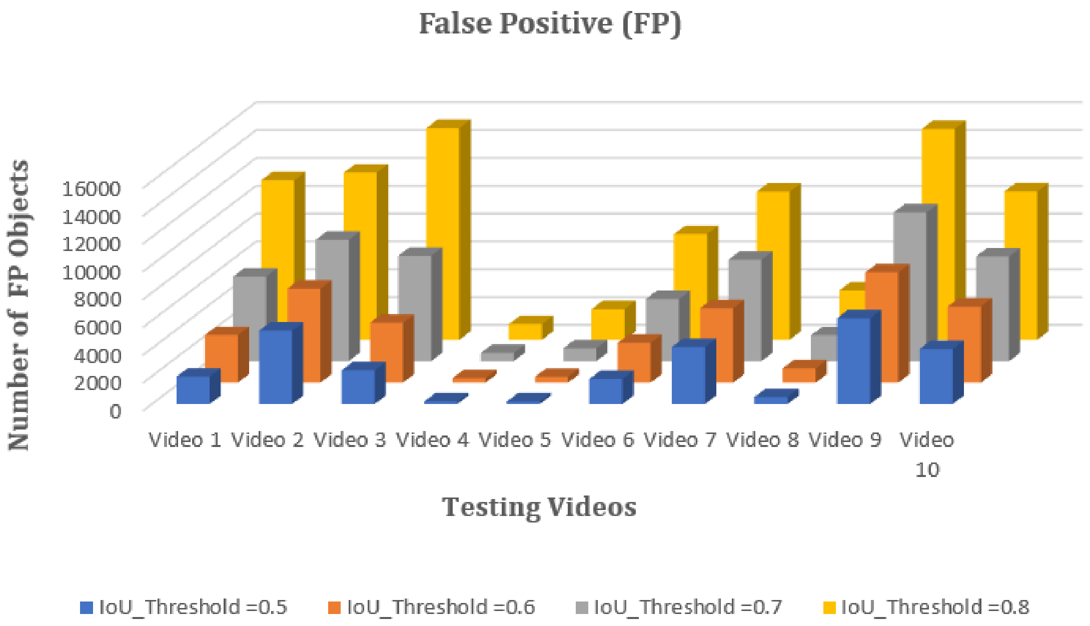

Figure 8. False Positive Values of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 8 illustrates number of False Positive (FP) objects of ten videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8) respectively. The highest of False Positive (FP) objects are video 3 and video 9 in the IoU-Threshold 0.8. These False Positive (FP) objects are over 14000 objects.

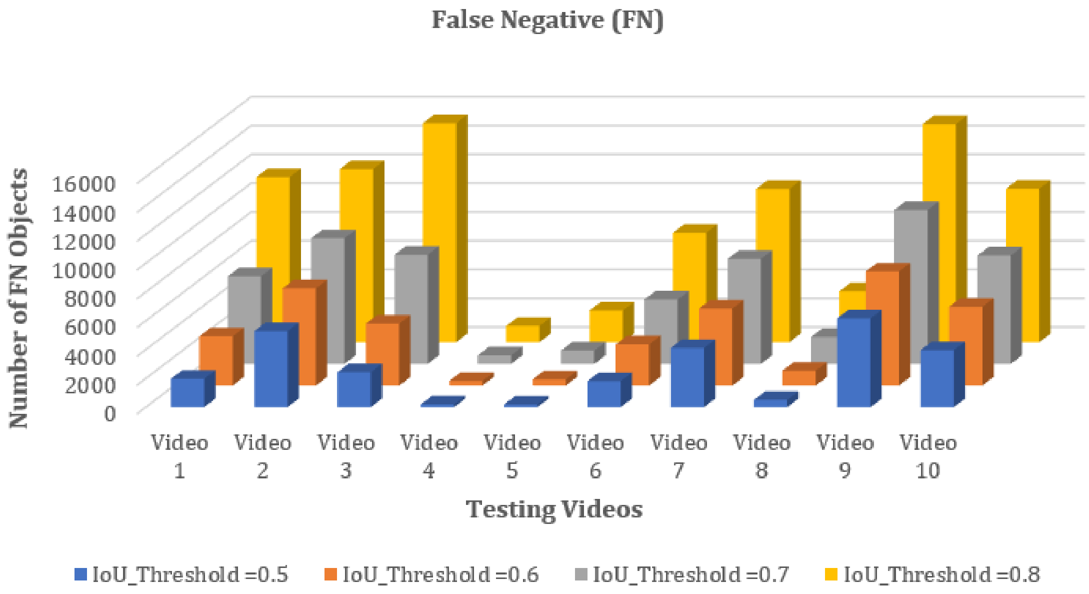

Figure 9. False Negative Values of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 9 describes number of False Negative (FN) objects of ten videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8) respectively. The highest of False Negative (FN) objects are video 3 and video 9 in the IoU-Threshold 0.8. These False Negative (FN) objects are also over 14000 objects.

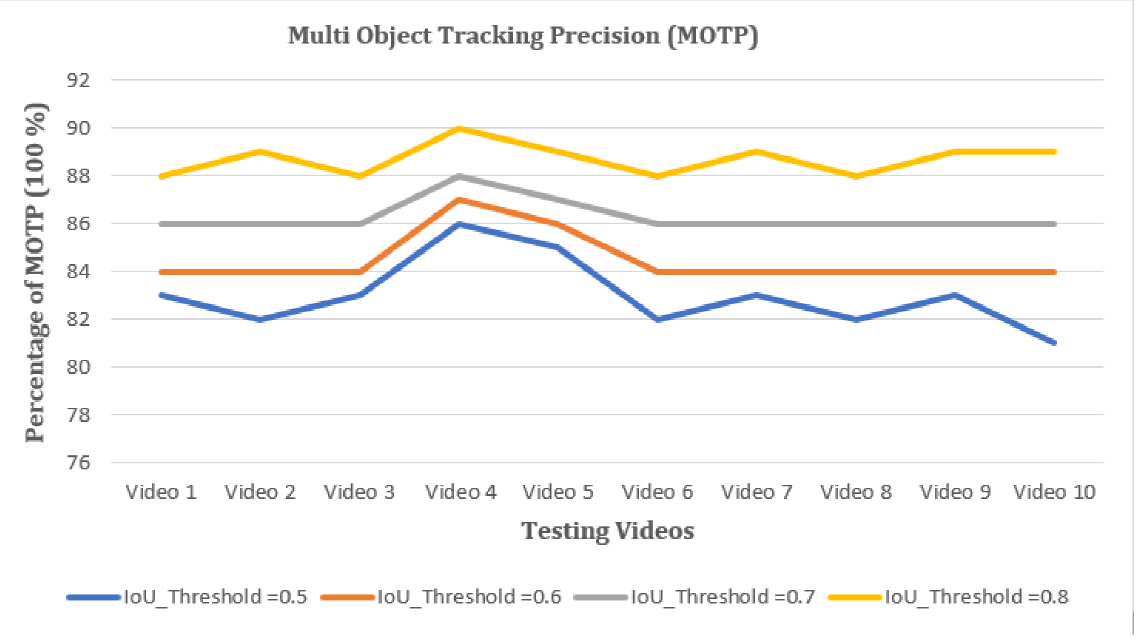

Figure 10. Multi Object Tracking Precision (MOTP) of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 10 shows Multi Object Tracking Precision (MOTP) values of ten videos. Among ten of videos, video 4 is the highest MOTP (90%) which is IoU-Threshold 0.8.

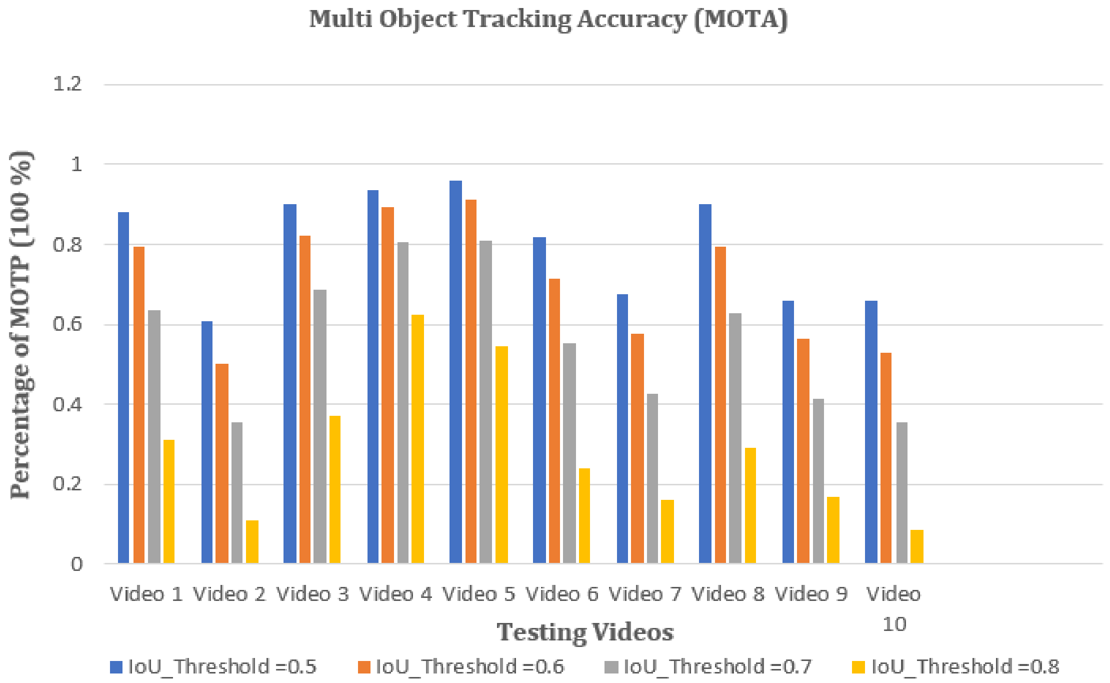

Figure 11. Multi Object Tracking Accuracy (MOTA) of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 11 illustrates Mutli Object Tracking Accuracy (MOTA) values of ten videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8) respectively. The highest Multi Object Tracking Accuracy (MOTA) is 96% at video 5 in IoU-Threshold 0.5.

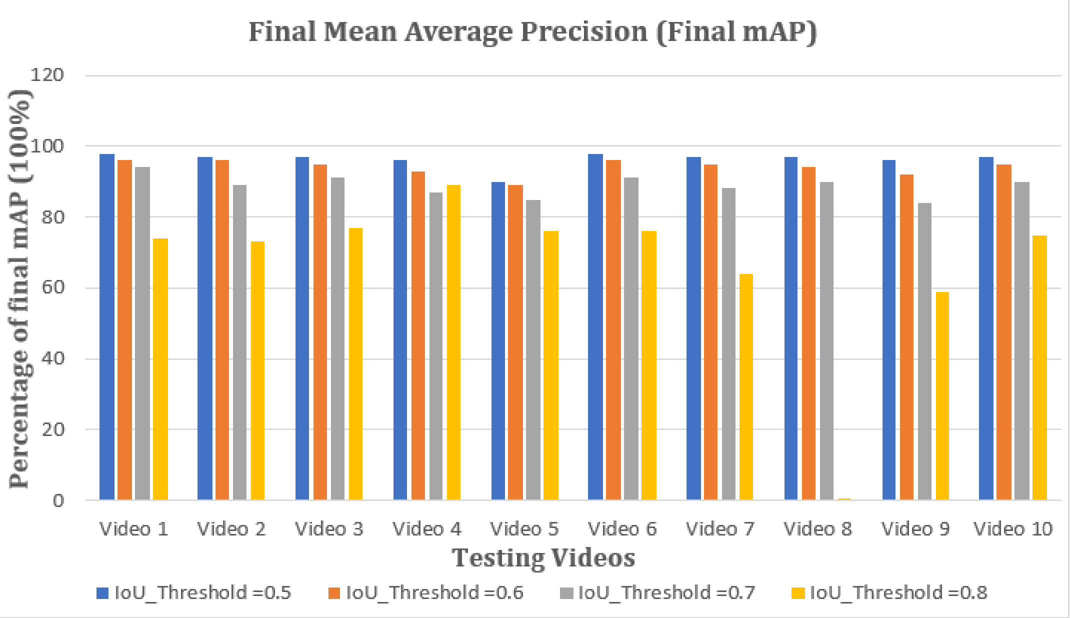

Figure 12. Final Mean Average Precision (mAP) of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8)

Figure 12 illustrates final Mean Average Precision (mAP) of Ten Videos depending on IoU-Threshold (0.5, 0.6, 0.7, 0.8). All of videos are the highest final mAP results at IoU-Threshold 0.5. Among them, video 1 is the highest percentage of Final mAP (98%).

D. Conclusion

In this paper, the proposed system is performed multi object detection with YOLOv8n and then the best detection results are obtained. After detection, multi objects tracking are achieved by using the hybrid method (FairMOT+stacked LSTM). Till now, there are no journal papers by using with the hybrid method for multi object tracking in football sport. According to the experimental results among ten videos, the average of the Multi Object Tracking Accuracy (MOTA) is 80 % at IoU-Threshold 0.5, the average of the Multi Object Tracking Precision (MOTP) is 89% at IoU-Threshold 0.8 and the average of the final mAP is 96% at IoU-Threshold 0.5 by using hybrid method for tracking These accomplishments are clear in the field of multi object detection and multi object tracking as a bright future.

E. Acknowledgment

I am grateful for the dedication, commitment, and professionalism demonstrated by my teachers. The constructive discussions, intellectual debates, and shared enthusiasm have fostered an enriching research environment. It has truly pushed the boundaries of our collective knowledge. This research stands as a testament to the power of collaboration and the collective pursuit of knowledge.

Finally I am deeply thankful to my univeristy Mandalay Technological University (MTU) for all satisfied achievements.

F. References

[1] B. Thulasya Naik and Md. Farukh Hashmi, "YOLOv3-SORT: detection and tracking player/ball in soccer sport" *Journal of Electronic Imaging* 011003-1, Vol. 32(1),Jan⁄Feb 2023.

[2] Jiating Jin, XingWei LI, Sha0jie Guan, "Multi-Object Tracking with Long-Short Term Memory," *ICIGP, Sanya*, China, January 01–03, 2021.

[3] Rekha B.S.l, Athiya Marium, Dr. G.N.Srinivasan3, Supreetha A. Shetty, "Literature Survey on Object Detection using YOLO"; *International Research Journal of Engineering and Technology (IRJET),* Volume 07, Issue 06, June. 2020.

[4] G.Jocher, A.Chaurasia and J. Qiu,"YOLO by Ultralytics", https://github.com/ ultralytics/2023.

[5] Yifu Zhang , Chunyu Wang , Xinggang Wang, Wenjun Zeng, Wenyu Liu, "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking", arXiv:2004.01888v5 [cs.CV] ,9 Sep 2020.

[6] Shipra Sahena, What is LSTM? Introduction to Long Short-Term Memory. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/#LSTM_Architecture.

[7] Manchester United, "Chido Double & Heaven Header _ Man Utd v Hong Kong, China," https://www.youtube.com/results?search_query=Chido+Double+%26+Heaven+Header+_+Man+Utd+v+Hong+Kong%2C+China.

[8] Shenzhen, Ultralytic YOLO vision. https://docs.ultralytics.com/models/yolov8.

[9] Medium.com,Introduction to tracker KPI,digital engineering centific. http://www.elsevier.com/authors.html, 1999, retrieved May 13, 2010.

[10] Shreyamainkar,https://www.kaggle.com/datasets/shreyamainkar/football-soccer.

[11] Nisma Novita Hasibuan, Muhammad Zarlis, Syahril Efendi "Detection and tracking different type of cars with YOLO model combination and deep sort algorithm based on computer vision of traffic controlling", Sinkron : Jurnal dan Penelitian Teknik Informatika,Volume 6, Number 1, October 2021.

[12] Heet Thakkar, Noopur Tambe, Sanjana Thamke, "Object Tracking by Detection using YOLO and SORT", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), CSEIT206256 , 2022.

[13] Muhammad Azhad bin Zuraimi, Fadhlan Hafizhelmi Kamaru Zaman, " Vehicle Detection and Tracking using YOLO and Deep SORT", 021 IEEE 11th IEEE Symposium on Computer Applications &amp; Industrial Electronics (ISCAIE), DOI: 10.1109/ISCAIE51753.9431784, Univ of Calif Santa Barbar, 2021.

[14] Shuman Guo, Shichang Wang, Zhenzhong Yang, Lijun Wang, Huawei Zhang,Pengyan Guo " A Review of Deep Learning-Based Visual Multi-Object Tracking Algorithms for Autonomous Driving," Appl. Sci, 12, 10741. https://doi.org/10.3390/app122110741, 2022.

[15] Thulasya Naik, Md.Farukh Hashmi,"YOLOv3-SORT: detection and tracking player/ball in soccer sport,"Journal-of-Electronic-Imaging, 15 Sep 2023.

[16] Fan Tang, Fang Yang, Xianqing Tian, "Long-Distance Person Detection Based on YOLO v7", Electronics , 12, 1502. https://doi.org/10.3390/2023.