ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2023.0412

OBJECT DETECTION IN HOCKEY SPORT VIDEO VIA PRETRAINED YOLOV3 BASED DEEP LEARNING MODEL

Suhas H. Patel[1] and Dipesh Kamdar[2]
[1]*Department of Electronics and Communication Engineering, Gujarat Technological University, India*
[2]*Department of Electronics and Communication Engineering, V.V.P. Engineering College, India*

Abstract

Object detection is the most common task in Sports Video Analysis. This task requires accurate object detection that can handle a variety of objects of different sizes that are partially occluded, have poor lighting, or are presented in complicated surroundings. Object in field sports includes player's team and ball detection; this is a difficult task resulting from the rapid movement of the player and speed of the object of concern. This paper proposes a pre-trained YOLOv3, deep learning-based object detection model. We have prepared a hockey dataset consisting of four main entities: Team 1 (AUS), Team 2 (BEL), Hockey Ball, and Umpire. We constructed own dataset because there are no existing field hockey datasets available. Experimental results indicate that the pre-trained YOLOV3 deep learning model generates comparative results on this dataset by modifying the hyperparameters of this pre-trained model.

Keywords:

Sport Video Analysis, Deep Learning, YOLOv1, YOLOv2, YOLOv3, Object Detection

1. INTRODUCTION

In recent years, computer science has shown promising potential in the sports industry. Compared to still photos, video footages cover more information about how situations change over time. Object detection from sport video needs more storage space and computational power than essential object detection. Still, it may be necessary for some circumstances where more than four objects are in a single frame[1]. Sports analysis is crucial for improving player performance. The use of computer vision-based virtual reality for posture correction in sports [2]. A computer vision-driven evaluation system provides support to make decisions for training in sports [3]. Application of object detection for sports analysis [4]. A traditional method uses sensors to locate and record athlete key positions. After that, data-driven suggestions are provided for training after the raw data has been analyzed using deep learning-based approaches [5]. Furthermore, adding more sensors will raise the price and it will affect adversely to athlete performance. Although due to tough competition one can wear sensors to identify their weakness and strengths.

It is hard for a coach to recall and analyze each player's motions and actions after a game to use that knowledge to instruct players and prevent potential blunders. As a result, a performance analyst, also known as a notational analyst, assumes the responsibility of documenting the entire event, gathering information about the player's activities, their movement, and the time of those activities, and then presenting those significant results to coaches [6]. However, the latest computer vision techniques, sport video analysis replace performance analysis. The primary function of this analysis technique is object detection from sport video frames by object detection various specific events from the sports are identified, which provide tremendous support to coaches for identifying player performance and team evaluation in particular condition [7].

To achieve objective of object detection from field hockey sport video the deep learning approach is implemented. In contrast to earlier machine learning methods, which required hand-crafting the features to be retrieved from the inputs, deep neural networks are able to learn and extract the features gained straight from the inputs [8].

The proposed model utilizes a pre-trained YOLOv3 based deep learning network which was precisely tuned for object detection from field hockey sport videos (Hockey player, Hockey ball, Umpire).

Since there isn't a publicly accessible standard hockey dataset for object detection, a dataset of hockey images is created from broadcasted gold medal hockey match of the Tokyo 2020 Olympics available on YouTube. The objects such as Players, Hockey ball, and Umpire are labeled in this hockey dataset and utilized for model training.

The primary contribution of this research is the automatic object detection model that is proposed for broadcasted hockey games using data from the author's datasets, which include the four primary hockey sports objects: AUS (Team 1), BEL (Team 2), Hockey ball, and Umpire.

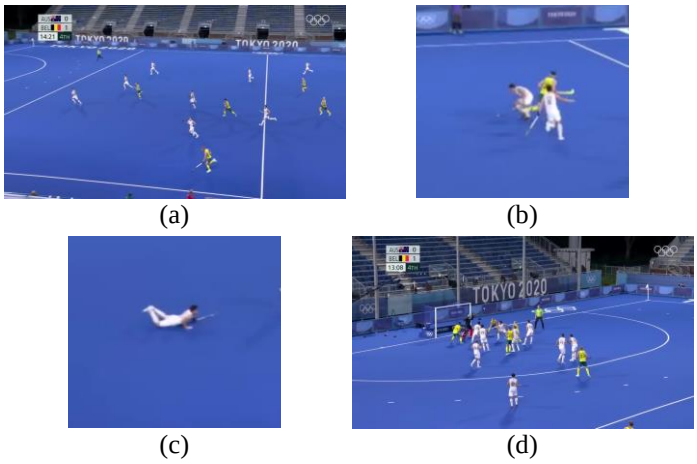
The rest of this paper is organized as given below; in Section 2 are reviews on related work for sports-related object detection methods. Section 3 focuses on methodology; Section 4 discusses experimental results and finally Section 5 is conclusion.

2. RELATED WORK

Sport analysis has become a popular research topic for many researchers due to the abundance of free internet datasets and the effective use of Convolution Neural Network(CNN) in object detection and image classification [9].

The traditional object detection model from sport video usually works to identify the players. It involves using different techniques such as connected component analysis[10], shallow convolutional neural networks [11], histogram of orientated gradients and support vector machines (HOG-SVM)[12], and deformable part model (DPM) [13]. The Fig.1 includes 4 images (a)-(d) that shows the various frames of sport video which targeted for object detection.

The Fig.1(a) is an example of a typical image where each object in the frames is separate from the others. Traditional model of object detection can identify the objects from these types of images, while it can barely detect objects in case of occlusion, self-occlude and situation as in Fig.1(b)-(d).



(a) (b) (c) (d)

Fig.1. (a) a normal hockey match condition (b) players position causes occlusion (c) one player lay down on ground. (d) Players are close to each other

Each sport has distinct characteristics, creating sport analysis a vast field. Before the sequence of deep learning development, scientists design each attribute to extract the desired features for a particular game in the early stages. Furthermore, the traditional object detection model includes identifying low-level features, which shows the position of the primary object, such as a soccer ball [14]. Pitcher style detection from baseball matches video use techniques such as an object segmentation algorithm [15].

Non-maximum suppression restricts the performance of the deformable part model (DPM) when detecting close players [16]. Other approaches, such as motion, pixel, or template-based methods, have various restrictions, such as the player should not stand still and cannot wear a white jersey, shorts, and socks [17].

Neural networks dominate object detection methods due to the rapid advancement of computer vision [18]. The YOLO object identification algorithms have improved from version 1 to version 6, outperforming conventional algorithms in capability and performance.

Deep learning-based approaches begin with text classification and progress over time to recognize human behavior, from a single action to a complicated group activity. Then it moves on to analyze sporting events in videos. Sport video analysis falls under several categories, including trajectory and tracking, applying spatial and temporal data, and combining spatial and temporal features.

In this work, we proposed a simple but efficient YOLOv3-based pre-trained deep learning model for object detection from hockey sport video.

3. METHODOLOGY

3.1 DATASET PREPARATION

As, there is no publicly accessible object detection dataset of hockey match. We use a self-prepared dataset for object detection from a YouTube video of a field hockey match between Australia and Belgium (Tokyo Olympics 2020 gold medal match), where four object names are AUS (Team 1), BEL (Team 2), Hockey ball, and Umpire. The video resolution was 1920×1080, and it splits into shorter duration video where the camera angle is in wide mode. These video clips are converted into frames that are manually annotated.

Total 1119 frames with 1920×1080 resolutions were manually labeled with four labels AUS (Team 1), BEL (Team 2), Hockey ball, and Umpire. Results are below optimum when model training is started right after data collection. Even if there are no issues with the data, implementing image augmentation increases the dataset and minimizes over fitting. Like tabular data, image data cleaning and augmentation can enhance model performance more than model architecture modifications. So here, in this case, images are preprocessed by image resize(640×640) and auto orientation, also augmented by rotation(-15 degree to +15 degree) and blur effect(up to 10 pix) as per shown in Fig.2, after Preprocessing and augmentation total of 2683 images as per details in Table.1 [19].

We can tell from our dataset that there is a significant class imbalance. Hockey Ball and Umpire are substantially less represented in our dataset than AUS and BEL which could interfere with the training of our model.

Table.1. Dataset with Pre-processing and without Preprocessing

	Without Pre-processing	**With Pre-processing and Augmentation**
Total Images	1119	2683
Classes	4	4
Unannotated	0	0
Training Set	783 (70%)	2347 (87%)
Validation Set	224 (20%)	224 (8%)
Testing Set	112 (10%)	112 (4%)
Annotation	12,937 (11.6 per Image (Average)	30939 (11.53 per Image (Average)
Average Image Size	2.07 mp	33.02 k
Median Image Ration	1920x1080	640x640
Class Instances		
1. AUS	5559(42.96%)	13293(42.96%)
2. BEL	5973(46.16%)	14258(46.08%)
3. Hockey_Ball	865(6.68%)	2075(6.70%)
4.Umpire	540(4.17%)	1313(4.24%)

Class imbalance in this dataset is due to no. of objects available in a single frame, as players are more in the single frame than the hockey ball and Umpire. In this case, the players have the highest priority for object detection, and it has almost equal representation in class instances.

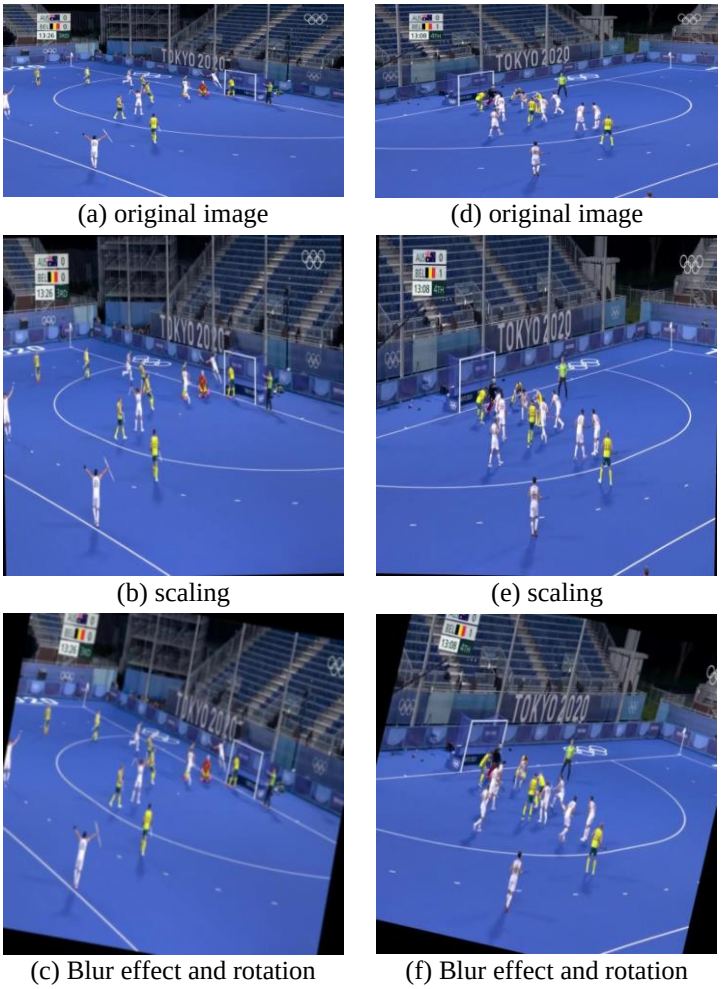
This dataset is used to train and evaluate deep learning models based on YOLOv3[20]. We use various clips of this match to test the effectiveness of our model.

3.2 OBJECT DETECTION MODEL

One of the most effective object identification techniques that developed from YOLO[21] and YOLOv2[22] is YOLOv3[20]. This specific model is a one-shot learner, meaning each image

only passes through the network once to make a prediction. The architecture is very performant, viewing up to 60 frames per second in predicting against video feeds. Fundamentally, YOLO is a convolutional neural network (CNN) that divides an image into subcomponents ($S \times S$ grid) and conducts convolutions on each of those subcomponents before pooling back to create a prediction. Each cell predicts B bounding boxes along with the confidence of these boxes. The confidence can reflect whether an object exists in the grid cell and, if it does, the IoU of the ground truth (GT) and predictions. YOLOv3 predicts objects in three different scales. Large things are detected by the 13×13 layer, whereas smaller objects are found by the 52×52 layer, and medium objects are found by the 26×26 layer. This can remedy the object size variation problem.

The Fig.4 shows the data flow of object detection model, where data collection and data preparation are essential steps based on how the model training performance is defined. Even though we're training our model on a custom dataset, instead of starting from scratch is advantageous to use another already trained model's weights as a starting point. Here the COCO dataset based pertained model is used for training our dataset.



(a) original image (d) original image

(b) scaling (e) scaling

(c) Blur effect and rotation (f) Blur effect and rotation

Fig.2. Some examples of image augmentation operations: original images, (top): scaling (middle), and Blur effect and rotation (down)

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig.3. Darknet-53 Model [20]

Fig.4. Object Detection Model Dataflow

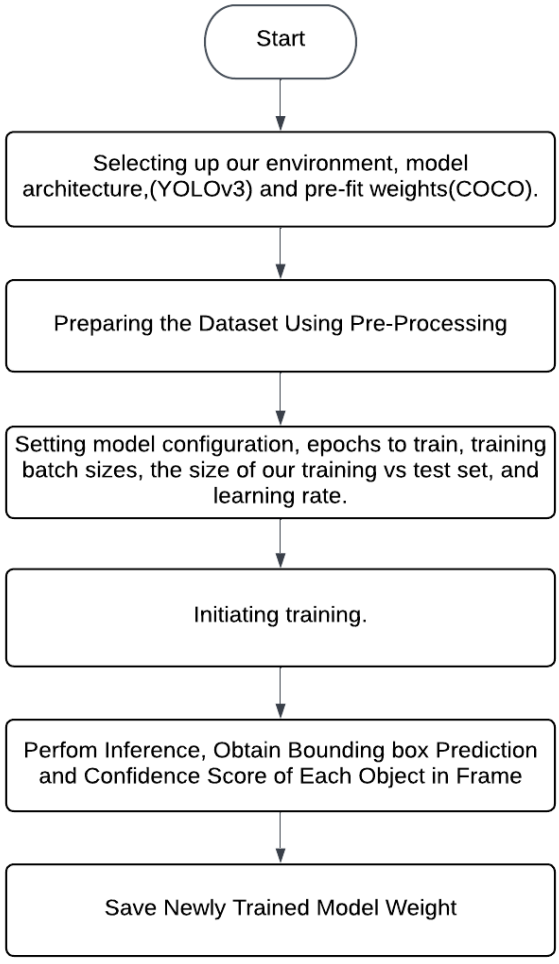


Fig.5. A flowchart of training and detection process of pretrained network YOLOv3

4. RESULTS AND DISCUSSION

This study conducted experiments on a Python 3 Google Compute Engine backend (GPU) using Google collaboratory. In this object detection system, YOLOv3-based model Darknet-53 has been pre-trained by COCO Dataset, as represented in Fig.3. The model receives images of 640x640 pixels as inputs; the batch size is set to 16. The model trained for 100, 200, and 300 epochs. The Scaled weight_decay is 0.0005, and Stochastic Gradient Descent (SGD) optimizer with parameter groups 72 weight, 75 weights (no decay) and 75 biases are set for this model. The video frames of hockey game are the input of the model. The inputs were passed on to the yolov3 model, fine-tuned for this hockey object detection task. The output was obtained from the highest confidence score of the bounding box after non-maximum suppression.

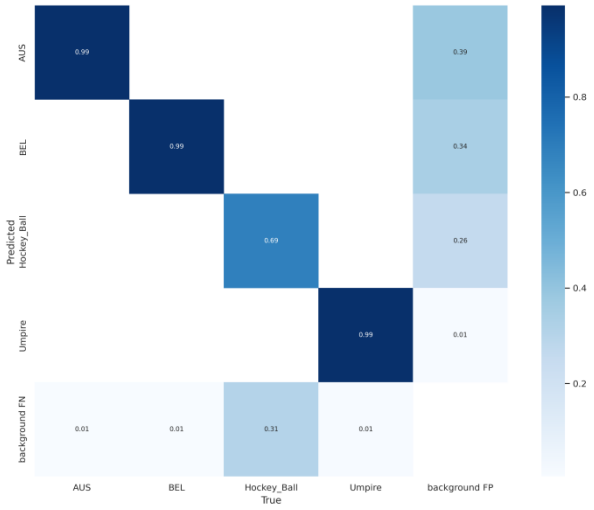
This model takes the complete visual frame and extracts features at the frame level. The YOLOv3 model passes through successive convolution layers from the first input layer to the last layer, learning patterns from the entire frame and extracting low-level features to high-level features.

This object detection model was implemented using pytorch (version 2.3.1) with annotation format of YOLO Darknet TXT. The training process was repeated three times by using different number of epoch 100, 200 and 300 to study one of the hyperparameter which is epoch. The batch size was fixed to the default value of 16. The models were trained with GPU provided by Google Colab. The framework of proposed model is illustrated in Fig.3.

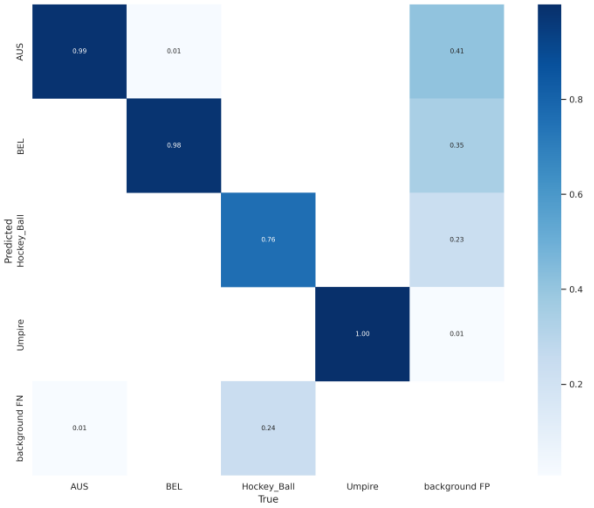
Table.2. The detection performance with different range of epochs (100 to 300).

No. of epochs	Class	Precision	Recall	F-1 score	mAP@ 0.5	Overall Accuracy (Map@0.5)
100	Aus	0.976	0.99	0.983	99.1%	88.9%
	Bel	0.98	0.99	0.985	99.3%	
	Hockey Ball	0.779	0.596	0.675	58.3%	
	Umpire	0.974	0.989	0.981	98.9%	
200	Aus	0.97	0.98	0.975	99%	91.2%
	Bel	0.99	0.994	0.992	99.4%	
	Hockey Ball	0.86	0.637	0.732	66.7%	
	Umpire	0.98	1	0.990	99.5%	
300	Aus	0.972	0.988	0.980	99%	91.3%
	Bel	0.988	0.992	0.990	99.4%	
	Hockey Ball	0.803	0.69	0.742	67.5%	
	Umpire	0.983	1	0.991	99.5%	

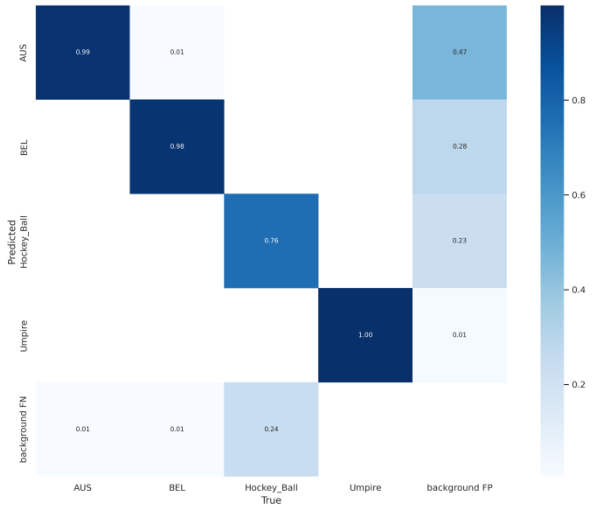
In this model, only a single hyperparameter is tuned, which is the number of epochs. This research is repeated with the same datasets and model architecture but with distinct epochs of 100, 200, and 300.



(a)



(b)

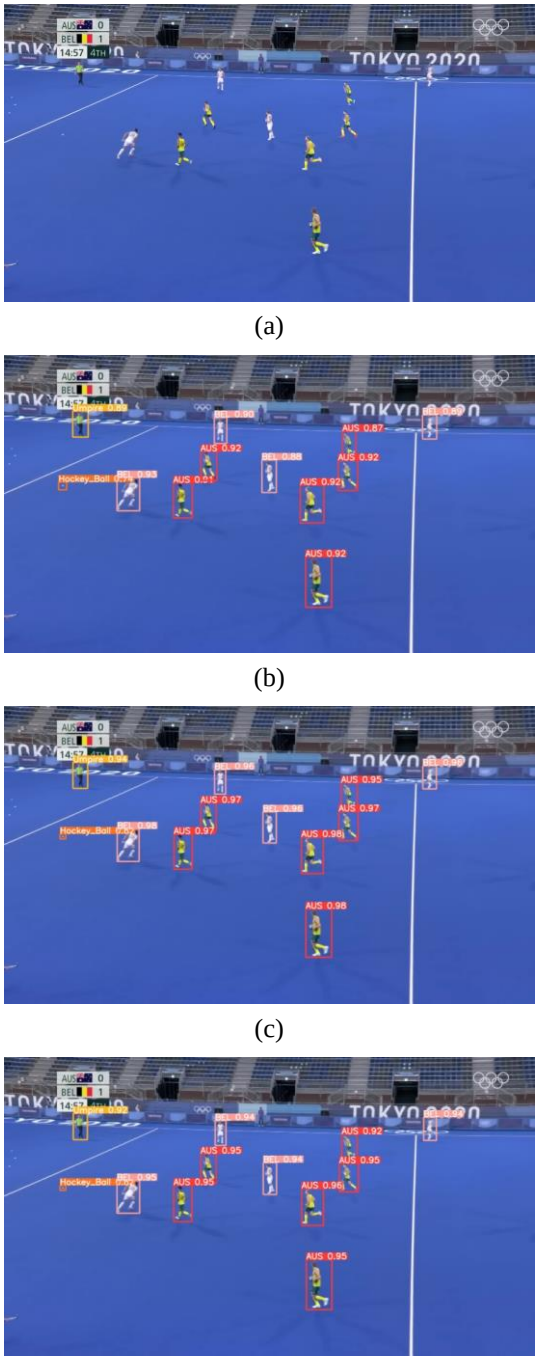


(c)

Fig.6. Confusion matrix of proposed method for different Numbers of epoch (a) 100, (b) 200 and (c) 300

The precision, recall, and F1 score, as well as confusion matrix for this research for each epoch, were presented in Table.2 and

Fig.6 respectively. Based on Table.2, we tabulated the precision, recall and F1 score of the model for each number of epochs.



(a)

(b)

(c)

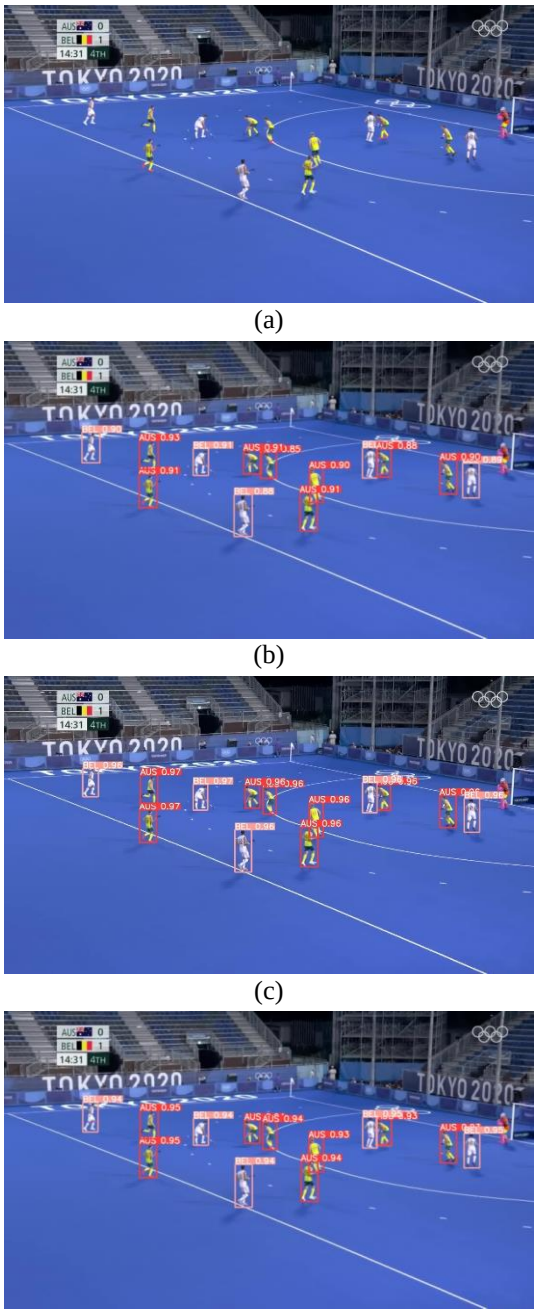
(d)

Fig.7. Output of object Detection model for first scenario (a) Input Image, Object Detection output for (b) 100 epoch (c) 200 epochs (d) 300 epochs

The F1 score assesses the balance between precision and recall, with precision measuring how precise a model is and recall measuring how many of the actual positives were properly predicted. We use the F1 score for the model's evaluation since, in the research, both precision and recalls are crucial for accuracy measurement. Based on the F1 score of the model, epochs with 300 have the highest score.

It has the highest accuracy as well which is 91.3%. The model was slightly confusing in Hockey Ball detection as it is a tiny object with a limited number of instances in the dataset. Fig.6 displays the confusion matrix used to evaluate the model's performance in terms of simple visualization. It is verified from confusion matrix that with higher number of epochs, model efficiency improved for object detection.

The Fig.7 and Fig.8 clearly show that the object detection model performance improved with increasing the number of epochs during training. It shows that a higher number of epochs provide the highest accuracy for object detection in the present scenario.



(a)

(b)

(c)

(d)

Fig.8. The output of object Detection model for second scenario (a) Input Image, Object Detection output for (b) 100 epochs, (c) 200 epochs, (d) 300 epochs

5. CONCLUSION

In this work, a deep learning-based transfer learning model, YOLOv3, has been proposed for object detection in field hockey. Four main objects AUS (Team 1), BEL (Team 2), Hockey ball, and Umpire, are identified from the collected hockey dataset through this pre-trained model. The highest accuracy achieved by the model is 91.3% in hockey object detection.

This paper provides a promising base for further research of event detection from field hockey sports. By selecting the optimal solution and adjusting it to a specific event, it could be possible to identify the player and the ball's location, such as the goal during the match.

REFERENCES

[1] Y. Wang, J.F. Doherty and R.E. Van Dyck, "Moving Object Tracking in Video", *Proceedings of International Conference on Applied Image Pattern Recognition*, pp. 95-101, 2000.

[2] C. Zhu, R. Shao, X. Zhang, S. Gao and B. Li, "Application of Virtual Reality Based on Computer Vision in Sports Posture Correction", *Wireless Communication and Mobile Computing*, Vol. 2022, pp. 1-15, 2022.

[3] L. Zhu, "Computer Vision-Driven Evaluation System for Assisted Decision-Making in Sports Training", *Wireless Communication and Mobile Computing*, Vol. 2021, pp. 1-15, 2021.

[4] M. Buric, M. Pobar, and M. Ivasic-Kos, "Object Detection in Sports Videos", *Proceedings of International Convention on Information and Communication Technology, Electronics and Microelectronics*, pp. 1034-1039, 2018.

[5] P. Salvo, A. Pingitore, A. Barbini and F. Di Francesco, "A Wearable Sweat Rate Sensor to Monitor the Athletes' Performance During Training", *Scientific Sports*, Vol. 33, No. 2, pp. 51-58, 2018.

[6] M. Stein, "Bring It to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 24, No. 1, pp. 13-22, 2018.

[7] I. McKeown, K. Taylor-McKeown, C. Woods, and N. Ball, "Athletic Ability Assessment: A Movement Assessment Protocol for Athletes", *International Journal of Sports Physical Therapy*, Vol. 9, No. 7, pp. 862-873, 2014.

[8] K. Rangasamy, M. A. As'ari, N. A. Rahmad, N. F. Ghazali and S. Ismail, "Deep learning in sport video analysis: a review", *Telecommunication Computer and Electronics Control*, Vol. 18, No. 4, pp. 1926-1946, 2020.

[9] G. Yao, T. Lei and J. Zhong, "A Review of Convolutional-Neural-Network-based Action Recognition," *Pattern Recognition Letters*, Vol. 118, pp. 14-22, 2019.

[10] R.G. Abbott and L.R. Williams, "Multiple Target Tracking with Lazy Background Subtraction and Connected Components Analysis", *Machine Vision and Applications*, Vol. 20, No. 2, pp. 93-101, 2009.

[11] A. Lehuger, "A Robust Method for Automatic Player Detection in Sport Videos 2 System Architecture 1 Introduction 3 Training Methodology 4 Player Localization", *Analysis*, Vol. 34, No. 1, pp. 1-14, 2007.

[12] S. Mackowiak, M. Kurc, J. Konieczny and P. Mackowiak, "A Complex System for Football Player Detection in Broadcasted Video", *Proceedings of International Conference on Signals and Electronics Systems*, pp. 119-122, 2010.

[13] D. Zhang, "Vehicle Target Detection Methods based on Color Fusion Deformable Part Model", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2018, No. 1, pp. 1-5, 2018.

[14] V. Pallavi, J. Mukherjee, A.K. Majumdar and S. Sural, "Ball Detection from Broadcast Soccer Videos using Static and Dynamic Features", *Journal of Visual Communication and Image Representation*, Vol. 19, No. 7, pp. 426-436, 2008.

[15] M. Leo, P. L. Mazzeo, M. Nitti and P. Spagnolo, "Accurate Ball Detection in Soccer Images using Probabilistic Analysis of Salient Regions", *Machine Vision and Applications*, Vol. 24, No. 8, pp. 1561-1574, 2013.

[16] Wei-Lwun Lu, J.A. Ting, J.J. Little and K.P. Murphy, "Learning to Track and Identify Players from Broadcast Sports Videos", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 7, pp. 1704-1716, 2013.

[17] M. Manafifard, H. Ebadi and H. Abrishami Moghaddam, "A Survey on Player Tracking in Soccer Videos", *Computer Vision and Image Understanding*, Vol. 159, pp. 19-46, 2017.

[18] A. Dhillon and G.K. Verma, "Convolutional Neural Network: A Review of Models, Methodologies and Applications to Object Detection", *Progress in Artificial Intelligence*, Vol. 9, No. 2, pp. 85-112, 2020.

[19] B. Dwyer and J. Nelson, "Roboflow (Version 1.0) [Software]", Available at https://roboflow.com, Accessed at 2022.

[20] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement", Available at https://pjreddie.com/media/files/papers/YOLOv3.pdf , Accessed at 2018.

[21] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", *Proceedings of International Conference on Pattern Recognition*, pp. 779-788, 2016.

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger", *Proceedings of International Conference on Pattern Recognition*, pp. 6517-6525, 2017.