YOLOv5 for Stroke Detection and Classification in Table Tennis

Bhuvana J, T.T. Mirnalinee, B. Bharathi,
Jayasooryan S, Lokesh N N
Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
(bhuvanaj,mirnalineett,bharathib)@ssn.edu.in
(jayasooryan19042,lokesh19055)@cse.ssn.edu.in

ABSTRACT

Sports action detection and classification is one of the most researched topics in video analytics. It is very useful in order to make fine tuned athletic training and get a better analysis of the athlete's performance. We present a model to detect and classify table tennis strokes made by players as a part of the MediaEval 2021 benchmark. Our approach extracts features using a YOLOv5 model trained on the MediaEval Fine Grained Action Detection and Classification of Table Tennis Strokes dataset provided to us, to detect and classify the moves/actions made.

1 INTRODUCTION

Action recognition is the task where predefined set of actions will be associated with the video. An automatic analysis of actions in the videos is the need of the day. In this paper we have proposed a method to detect and classify strokes in a dataset consisting of various strokes in table tennis performed during a match or during practice. Localization of the objects and identifying them followed the classification is the sequence of tasks involved in the action recognition. Strategic decisions can be taken once the actions are detected and classified. The dataset consists of 20 different classes [5] of strokes which the detection and classification is based upon, and these moves are shot in natural conditions. Application of machine learning in this specific domain can improve athletic performance by computer-aided analysis of moves. We implemented a YOLOv5 model which is based on CNNs for this problem and discussed our results with the given dataset.

2 RELATED WORK

Sports action classification is a topic in which there has been a lot of research been carried out which tend to focus on recognising a large number of actions using spatio-temporal models, using videos. Feature extraction, dictionary learning , and classification [7] are the steps involved in Action localization and recognition of sports videos. Sliding window approach is used to choose the maximum score of the classifier in the spatio-temporal volume. Siamese Spatio-Temporal Convolutional Neural Network (SSTCNN) has been used to detect the table tennis strokes [6]. It uses the RGB video frames and Optical Flow normalization to enhance their performance. Similar action recognition research has been found in literature [3], [1] using 3D ConvNets and extracting HOG of the Temporal Difference Map (TDMap) respectively. Long-term Recurrent Convolutional Network (LRCN) has been used to classify the table tennis strokes [8] that extracts the features using VGG16, a pretrained model. Our approach does not use optical flow data to detect the moves and instead directly uses the frame sequences.

MediaEval'21, December 13-15 2021, Online

3 APPROACH

The dataset had many videos which consisted of actions and moves made by players which had very subtle differences amongst them. So we took into account temporal information in the frames in an effective manner. Since the actions had very subtle differences with the low inter-class variability, it was a difficult task to handle. CNN models found to be optimal to classify if the data is highly spatial with proper discrimination among the classes. We decided to study this with object detection and recognition deep learning framework, YOLOv5 architecture [2].

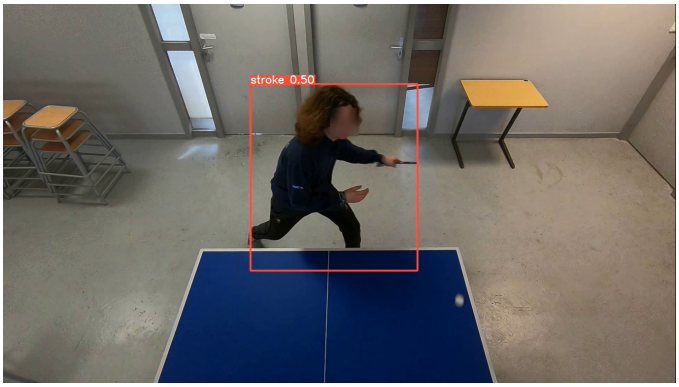
3.1 Data Pre-processing

The YOLO model takes fixed input sizes for each mini batch. The frames were downscaled to 512×512 in order to keep the size of the files manageable. CVAT (Computer Vision Annotation Tool) was used to annotate the actions of the players, by drawing bounding boxes over the body, focused on the hand holding the bat, in the videos as per the given frame number in the dataset. The strokes annotated are of varying duration with some being very short while others more lengthy. This meant we had to ensure the extracted frames had information on the entire move in it, irrespective of the duration. Since there were two different annotation data sets i.e training and classification, we observed that the detection frame sequences were overlapping with the classification ones, we annotated only using the detection data set using the stroke-classes when present, or marked it as just "stroke" if no class was present. We then split the annotated files in the required two file types. This saved us a lot of work, as we did not have to annotate the same video twice or draw two bounding boxes.

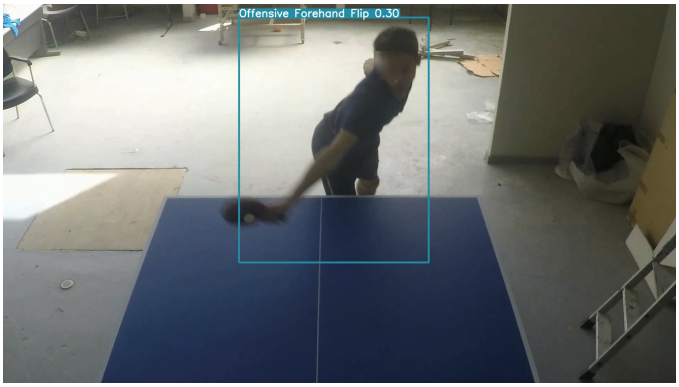
3.2 Proposed Model

Our approach uses the complete RGB frame sequence of the whole video that consisted of some subset. The YOLOv5 architecture is a modified version of the YOLOV4 implemented by Ultralytics. YOLOv5 has three functional components namely the Backbone, CSPDarknet, its Neck, PANet and the Head, Yolo Layer. CSPDarknet helps to extract the features from the frames of the table tennis videos. Feature pyramids are constructed using the PANet stage that helps in generalizing with different sized objects. By applying the bounding boxes on the features the head layer performs the object detection task.

The YOLOv5 model has been trained for 15 epochs in order to detect the strokes, classify them and find their respective bounding



(a) Detected as stroke



(b) Detected as offensive forehand flip

Figure 1: Stroke detections

boxes from the frames. The model is trained to detect 20 different classes of strokes. It has obtained a training and validation loss of about 0.0039% and 0.0021% with loss function mentioned in Table 1. Hyper-parameters adopted by our approach is listed in Table 1. As we considered the whole action sequence during detection, over-fitting was a major problem as even a still position of the player was fed to the model with a positive label which caused over-fitting of the model. This could have been prevented by taking only the frame sequences where a move was performed. In the classification part, this model seemed to perform better than the detection part as the classes were in lesser amount in the dataset compared to the detection part.

4 RESULTS

The model was able to classify 13 out of the 20 classes as we could not annotate the videos which had the other 7 classes. We achieved an accuracy of 9.95% on the 13 classes where some of the classes were predicted with a good accuracy and some of them had poor accuracy. It has been observed with respect to per class accuracy that the model learnt some of the moves well than the others. A sample images after detection are shown in Figure 1a and 1b.

But a very poor performance in the test set of the detection (mAP=0.000525 G-IoU=0.247) showing that using YOLOv5 with current training was not a probable model for this dataset. The G-IoU is better than mAP shows that the detection is moving towards the ground truth, that if we have trained the network with different hyper-parameters for more epochs the detection would have been better. The baseline results for this dataset can be seen in [4] Hyper-parameters adopted by our approach is listed in Table 1. The model was not able to completely detect the moves such as serve backhand topspin, serve backhand backspin, forehand loop and forehand side spin leading to very poor test accuracy. The model did not perform well on the detection part as the frame sequences not only moves but other actions as well such as standing still, walking, etc. This resulted in incorrect detection of stroke from the frame sequences. A closer analysis shows that the model fails to distinguish between the moves belonging to a specific class (such as Serve, Defensive, Offensive) as the differences are very intricate. The model tended to prefer certain moves significantly more than others on the test set, which arose due to the distribution of the training set. Using uniform amounts of data to work with resulted in the number of examples to train on being very low. The difference in accuracy on test and validation data might be due to the frequency of the different classes on the test set being different from the training and validation set.

Table 1: Hyper-parameters used

Hyper-parameter	**Value**
Learning Rate	0.01
optimizer	adam
Loss	Binary Cross-Entropy with Logits Loss
Momentum	0.937
Weight Decay	0.0005
IoU Threshold	0.2
Anchor Threshold	4.0

5 DISCUSSION AND OUTLOOK

As we processed the data where each move/action was considered for a very large frame sequence, it resulted in over-fitting. Thus over-fitting could have been avoided if the moves/actions were precisely annotated in the video dataset and considered correctly when fed into the model. We learnt that data of this kind needs to have precise annotations after pre-processing which could result in better results. Thus our model could not show comparable accuracy when compared with the baseline model which was provided for reference. The performance could have been enhanced further by annotating all the videos and by training for more number of epochs. Conv3d with different hyper-parameters other than the baseline model can be attempted to study the performance.

REFERENCES

[1] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, Ahmed Bouridane, and Azeddine Beghdadi. 2021. A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence* 51, 2 (2021), 690–712.

[2] Glenn Jocher. 2020. YOLOV5. https://github.com/ultralytics/yolov5. (2020). Online; accessed 29 October 2021.
[3] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 221–231.
[4] Pierre-Etienne Martin. 2021. Spatio-Temporal CNN baseline method for the Sports Video Task of MediaEval 2021 benchmark. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.
[5] Pierre-Etienne Martin, Jordan Calandre, Boris Mansencal, Jenny Benois-Pineau, Renaud Péteri, Laurent Mascarilla, and Julien Morlier. 2021. Sports Video: Fine-Grained Action Detection and Classification of Table Tennis Strokes from videos for MediaEval 2021. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.
[6] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. *Multimedia Tools and Applications* 79, 27 (2020), 20429–20447.
[7] Khurram Soomro and Amir R Zamir. 2014. Action recognition in realistic sports videos. In *Computer vision in sports*. Springer, 181–208.
[8] Siddharth Sriraman, Srinath Srinivasan, Vishnu K Krishnan, J Bhuvana, and TT Mirnalinee. 2019. MediaEval 2019: LRCNs for Stroke Detection in Table Tennis.. In *MediaEval*.