IAIC International Conference Series
Vol. 4, No. 1, 2023, pp. 132 ~ 138
E-ISSN: 2774-5899 | P-ISSN: 2774-5880
DOI: https://doi.org/10.34306/conferenceseries.v4i1.643

Detecting and Tracking Player in Football Videos Using Two-Stage Mask R-CNN Approach

[1*]Amir Mahmud Husein, [2]Chalvin, [3]Kalvintirta Ciptady, [4]Raymond Suryadi, [5]Mawaddah Harahap
[1,2,4,5]Informatics Engineering, Prima Indonesia University, Indonesia
[3]Computer Science, Prima Indonesia University, Indonesia
E-mail: [1*]amirmahmud@unprimdn.ac.id, [5]mawaddah@unprimdn.ac.id
***Corresponding author**

(Received November 8, 2023 Revised November 17, 2023 Accepted November 25, 2023, Available online December 19, 2023)

Abstract

Football is one of the most popular sports worldwide and capable of attracting the attention of millions of fans to a single match in the top leagues. The English Premier League, Spanish LaLiga, German Bundesliga, Italian Serie A, and French Ligue 1 are the five best leagues in the world today. There was an experiment where researchers want to analyze the efficiency and accuracy percentage of tracking and detection using the deep learning method of the Mask R-CNN model in classifying positive and negative X-Ray images in football matches. In this study, we applied Mask R-CNN for the segmentation and detection of football players. This model was based on two different backbones, namely ResNet101 and DenseNet. Both backbones produced accuracy values that were not significantly different, but the DenseNet approach performed better than ResNet101 based on testing results in the validation and testing sets. Based on comprehensive experiment results on the dataset, it has been shown that the Mask R-CNN approach with DenseNet can achieve better results compared to Mask R-CNN with ResNet101. Due to insufficient understanding of the characteristics of image types and the uneven distribution of various types of data sourced from random videos, there was still room for improvement in the trained model.

Keywords: *Deep Learning, DenseNet, Football, Mask R-CNN, ResNet-101*

1. Introduction

Football is one of the most popular sports worldwide, capable of attracting the attention of millions of fans to a single match in the top leagues. The English Premier League, Spanish LaLiga, German Bundesliga, Italian Serie A, and French Ligue 1 are the five best leagues in the world today. In recent years, there has been an increased interest in using data to enhance analysis in football, given technological advances that make live broadcasts available to everyone [1].

With the explosive growth of sports video data on internet platforms, managing this information scientifically poses a significant challenge in the current era of big data. In recent years, deep learning has made significant progress in object detection and action detection research, but there have been few achievements in sports video detection [2]. [3] proposed a segmentation algorithm based on deep learning to detect, group players, and extract player spatial features with the aim of realizing player pose estimation. Uchida et al [4] proposed a new automated method to detect offside in football match videos. Not only in football, but [5] applied data analysis in Basketball match videos and [6] for Badminton sports.

Football is a beloved sport, and its broad viewership makes football videos one of the most valuable types of videos for analysis. Researchers have achieved certain results in football video content

analysis. Finding interesting clips from complete long videos is an urgent problem to address in football match video analysis. The details of sports event detection with traditional machine learning are relatively coarse, and the types of events that can be detected are limited. In recent years, advanced sports analysis has been adopted in most major sports leagues [7].

Player and game statistics generated from football match analysis serve various purposes. The results can assist coaches in improving team tactics as they provide valuable insights into player performance in specific situations or information about the overall fitness level of the players. Extracted data can also be used by fans who want to know as much as possible about their favorite players or teams. An automated approach that could be used for this task is by employing deep learning algorithms. These algorithms have the ability to learn from existing training data by adjusting their internal structure. The knowledge learned can then be applied to unseen data.

The development of deep learning algorithms in recent years has had a significant impact in various industries such as healthcare, filmmaking, marketing, sports, and more. Football match analysis in the last decade has attracted a lot of researchers' interest to develop various algorithms with different objectives. Mask R-CNN is a Convolutional Neural Network (CNN) and is state-of-the-art in image segmentation. This variant of the Deep Neural Network detects objects in images and produces high-quality segmentation masks for each instance [8].

Mask R-CNN was developed on top of Faster R-CNN, a Region-Based Convolutional Neural Network. Mask R-CNN is built using Faster R-CNN. While Faster R-CNN has 2 outputs for each candidate object, class label, and bounding box offset, Mask R-CNN is the addition of a third branch that outputs object masks [9]. The additional mask output is different from the class and box outputs, requiring much better spatial layout extraction of an object. Mask R-CNN is an extension of Faster R-CNN and works by adding a branch to predict object masks (Region of Interest) in parallel with the existing branch for bounding box recognition [10]. The key element of Mask R-CNN is pixel-to-pixel alignment, which is a crucial part of Fast/Faster R-CNN that was missing. Mask R-CNN adopts the same two-stage procedure as the first stage (i.e., RPN) [11]. In the second stage, in parallel to predicting class and box offsets, Mask R-CNN also outputs binary masks for each RoI. This differs from recent systems where classification relies on mask predictions [12]. Moreover, Mask R-CNN is easy to implement and train due to the Faster R-CNN framework, facilitating various flexible architecture designs. Additionally, the mask branch only adds a small computational overhead, allowing for fast systems and quick experiments [13].

In the context of this research, the object of study is football videos because compared to other sports games, football has a larger amount of data conducive to data analysis, a broader audience group, and rare content [14]. The research approach here includes semantic video segmentation and feature extraction by applying deep learning algorithm network rules by analyzing video frame capture rules. A two-stage deep learning algorithm is used to build the model [15]. The proposed model architecture on Mask R-CNN implements two approaches based on ResNet-101 and DenseNet backbones as a comparison in detecting players in video scenes [16].

2. Research Method

The model used in this research is Mask R-CNN model with the purpose of extracting and cleaning data that will be processed [17].

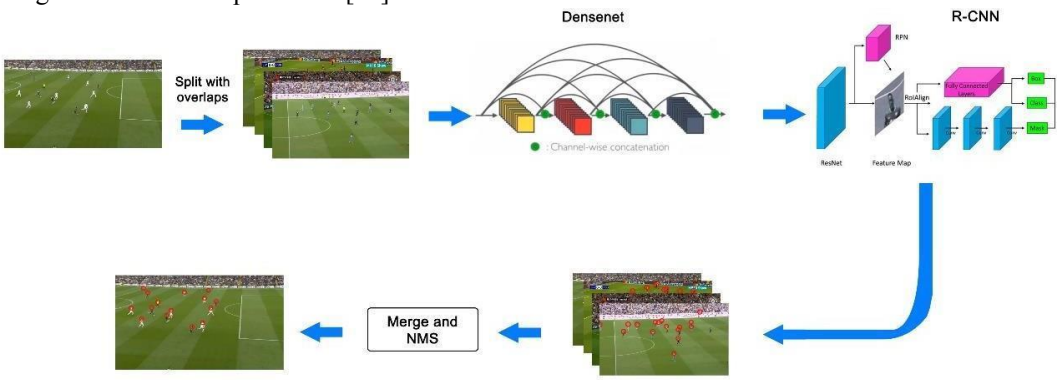


Figure 1. Our Method Framework

The framework for the testing conducted using football player detection on a dataset is explained in the following stages:

1. Searching for football highlight videos and saving them to a folder. The primary focus is the faces of football players captured by the camera.
2. The next step is utilizing deep learning for storing player facial data.
3. Then, employing the Mask R-CNN model to extract digits from machine bounding boxes.
4. After evaluation, two additional approaches will be tested in detecting the number of video frames and frame speed.

The dataset that we use in this research is a collection of football match highlight videos from various football leagues in Europe [18]. Each video is captured in snapshots and saved to a USB flash drive. Stored photos from the video snapshots are used as the primary data for analysis [19]. Figure 2. Shows samples of the video snapshots.

Figure 2. Samples of collected video snapshots

3. Results and Analysis

3.1. Data Preparation

The video data collection consists of match videos sourced from YouTube with a total of 32 English league match videos to be used as the training and testing dataset. The overall videos have an average match duration ranging from 1 minute to 2 minutes.

Training a deep learning algorithm based on the Mask R-CNN model requires a substantial amount of data. To avoid the impact of model overfitting, imbalanced data distribution, or using a single background in model training and testing, the first step is to augment the data to expand the dataset. For this purpose, we extracted the video dataset into several randomly chosen images frame by frame from each video during the match [20]. This process resulted in 9,055 images. The entire set of these image data is used as the dataset, then divided with a ratio of 6:3:1, comprising 5,054 images for the training set, 2,700 images for the validation set, and 900 images for the testing set.

3.2. Mask R-CNN-Dense Net

The backbone is the ConvNet architecture that will be used in the first step of Mask R-CNN. By default, Mask R-CNN has backbones such as ResNet50, ResNet101, and ResNext101. The choice between these should be based on the trade-off between training time and accuracy [21]. In this study, we propose the architecture of the DenseNet network as a comparison because this architecture has the advantage of narrower network layers and fewer parameters. This is mainly due to the design of densely connected blocks as illustrated in Figure 3.

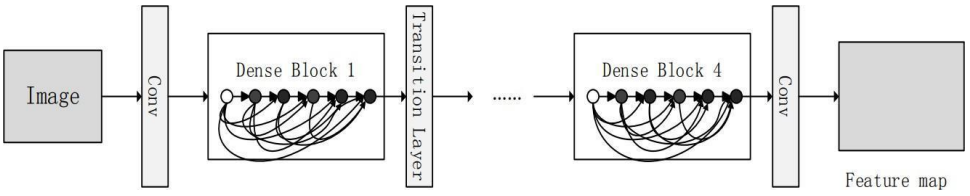


Figure 3. DenseNet architecture

DenseNet is utilized as the feature extraction backbone with the aim of obtaining various levels of feature maps [22]. To combine features of different scales from images, a feature pyramid network is employed to create dual-scale feature maps. Subsequently, the region proposal network is used to take the feature map of the image as input and output a set of rectangular object region proposals. Each region

proposal is accompanied by a score, representing the probability within the proposal region [23]. DenseNet can effectively acquire feature map extractions from images, which are then fed into the regional proposal network and three head networks [24]. DenseNet is a densely connected convolutional neural network model, often referred to as a composite network. In essence, DenseNet is a convolutional neural network containing one or more densely connected modules [25].

3.3. Training

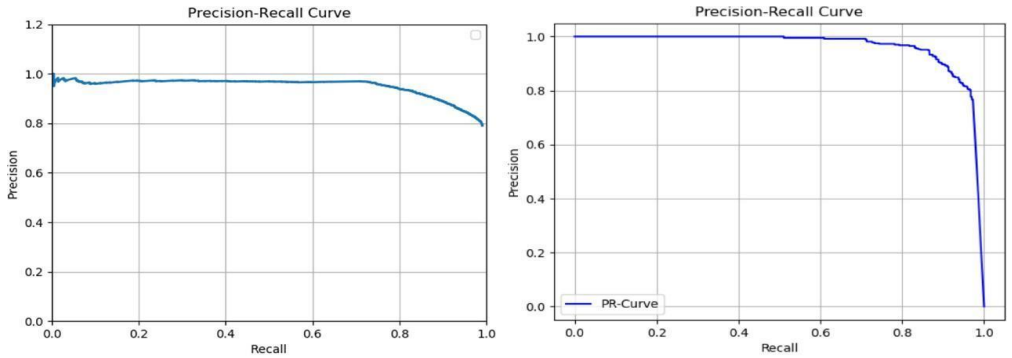
In this study, the training process is done within 80 epochs and each epoch is trained for 500 steps during the training period. Throughout the training process, as Mask R-CNN can only utilize three-channel RGB images for prediction, the channels from the test data set of 900 RBGA images are modified to RGB after an error was identified. Figure 4. depicts the loss diagram for each partial function in this research.

Figure 4. Training losses and validation losses of Mask R-CNN with (a) ResNet101 and (b) DenseNet101 as the backbones

In Figure 4, the results show the overall loss of the Mask R-CNN model based on ResNet101 and DenseNet101 backbones with 80 epochs where each epoch is trained with 500 steps. Comparative experiments are conducted on the same dataset at different learning rates. From the training results, when the learning rate is set to 0.001, the ResNet backbone results in a training loss value of 0.3099 and the validation loss that is decreased to 0.4637. Meanwhile, the DenseNet backbone's training loss is decreased to 0.0434 and its validation loss drops to 0.0601. These results indicate that both models are effective for this training.

3.4. Testing

The testing process is carried out based on weight files obtained from the training of Mask R-CNN and is used to evaluate the trained model. The remaining weight file from the last training iteration in the training process is selected to evaluate the test set for both models. Precision (P), Recall (R), Average Precision (AP), and Mean Average Precision (MAP) are used as the main parameters to evaluate the models in this study.



(a) (b)

Figure 5. Precision-Recall curves of Mask R-CNN with (a) ResNet101 and (b) DenseNet101 as the backbones

Figure 5. Shows the performance of the proposed Mask R-CNN model based on two different backbones, namely ResNet101 and DenseNet101. The performance of both models does not exhibit significant differences. Different performance tests on the test set using weights obtained from the training set after 80 epochs at different learning rates are plotted for Precision-Recall (PR) curve response at a learning rate of 0.001. Additionally, the same operational testing is conducted on the validation set with weights trained by the training set. Based on the testing results from both the validation and test sets for both models, it is found that Mask R-CNN with the ResNet101 backbone achieves 87.90% for the validation set and 87.52% for the test set. On the other hand, mAPs for the DenseNet backbone model yield 95.22% on the validation set and 99.45% on the test set, indicating that the proposed Mask R-CNN model with the DenseNet101 backbone performs better than the default Mask R-CNN with the ResNet101 backbone. The performance results of both models are shown in Table 1.

Table 1. MAP (%) of Mask R-CNN

Backbone	Validation Set	Testing Set
ResNet	87.90	87.52
DenseNet (proposed)	95.22	99.45

Based on the conducted testing, the next step involves testing a new video downloaded from YouTube to demonstrate the performance results of Mask R-CNN-DenseNet. The testing results for the video can be seen in the following image scenes.

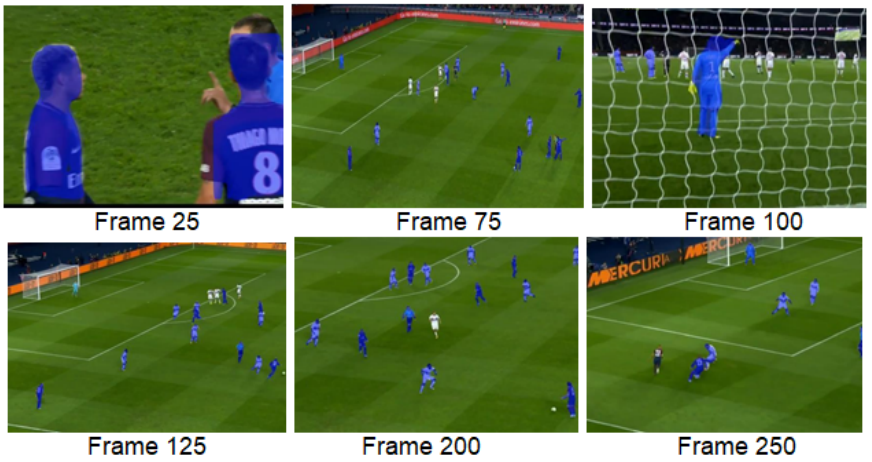


Figure 6. Testing results of Mask R-CNN-DenseNet on a new video

Figure 6. Displays the testing results for a football match video using the DenseNet101 backbone in the Mask R-CNN model. It is observed that there are still some video scenes where segmentation is not successful. The segmentation accuracy is significantly influenced by the distance of the video scenes. The

further the video frame, the lower the segmentation accuracy is, as shown in frames 75 and 125. This presents a challenge that needs further evaluation and will be our focus in future works.

3.5. Discussion

Mask R-CNN is developed on top of Faster R-CNN, a Region-Based Convolutional Neural Network built using Faster R-CNN. This model offers several backbone options for object segmentation, including ResNet50, ResNet101, and ResNext101. The choice between these options should be based on the trade-off between training time and accuracy. ResNet50 tends to require relatively less time compared to newer options and has some pre-trained weights available for large datasets like COCO, which can significantly reduce training time for various instance segmentation projects. ResNet101 and ResNext101, while requiring more training time due to the increased number of layers, tend to be more accurate if there are no pre-training weights involved, and key parameters such as learning rate and epochs are properly tuned. The ideal approach is to start with pre-trained weights such as COCO with ResNet50 and evaluate the model's performance. This approach works faster and better for models involving real-world object detection trained on the COCO dataset. If accuracy is crucial and high computational power is available, options like ResNet101 and ResNext101 can be explored.

In this study, Mask R-CNN is implemented for football player segmentation and detection. The model is based on two different backbones: ResNet101 and DenseNet101. Both backbones used yield accuracy values that are not significantly different, but the DenseNet approach performs better compared to ResNet101 as seen in the testing results on the validation and testing sets in Table 1. Additionally, the proposed Mask R-CNN-DenseNet model's performance is evaluated based on different findings from various studies as shown in Table 2.

Table 2. Comparison with Other Studies

Model	Accuracy (%)
TCN (14)	99.3
Faster R-CNN-DETR (15)	91.7
Faster R-CNN-ResNet (16)	98.0
Mask R-CNN-DenseNet (proposed)	99.4

Table 2. Presents a comparison with other researchers where the proposed results show better performance compared to other studies. However, this comparison focuses solely on the accuracy of the proposed model even though the testing datasets are different. Nevertheless, there are still limitations in the application of DenseNet in this research, such as a lack of understanding of image data sourced from videos, augmentation, and testing with different backbones like ResNet50, ResNext101, Xception, and others. Further research could consider preprocessing before training.

4. Conclusion

In this study, we propose the use of Mask R-CNN for the segmentation and detection of football players based on a dataset of football match videos from the English league sourced from YouTube. Different backbone implementations in the Mask R-CNN model are applied and tested based on the evaluation metrics of Precision (P), Recall (R), Average Precision (AP), and Mean Average Precision (mAP). The performance testing process for both models involves 80 epochs with 500 steps in each epoch. Based on the results of comprehensive experiments on the dataset, it has been demonstrated that the Mask R-CNN approach with DenseNet achieves better results compared to Mask R-CNN with ResNet-101. However, due to insufficient understanding of the characteristics of image types and the uneven distribution of various types of data sourced from random videos, there is still room for improvement in the trained model.

References

[1] Sun P, Zhao X, Zhao Y, Jia N, Cao D. Intelligent Optimization Algorithm of 3D Tracking Technology in Football Player Moving Image Analysis. Wirel Commun Mob Comput. 2022;2022(1).

[2] A. G. Prawiyogi, M. Hammet, and A. Williams, “Visualization Guides in the Understanding of Theoretical Material in Lectures,” *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, pp. 54–60, 2023.

[3] Radke D, Orchard A. Presenting Multiagent Challenges in Team Sports Analytics Blue Sky Ideas Track. Proc Int Jt Conf Auton Agents Multiagent Syst AAMAS. 2023;2023-May:1781–5.

[4] D. Iriani, S. Parman, A. F. Hafizh, I. Rachmawati, and Y. A. Solihah, “Ambient Media

Advertisement of Catur Insan Cendekia University to Improve Brand Awareness," *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 97–110, 2023.
[5] Liu N, Liu L, Sun Z. Football Game Video Analysis Method with Deep Learning. Comput Intell Neurosci. 2022;2022.
[6] K. Diantoro, D. Supriyanti, Y. P. A. Sanjaya, and S. Watini, "Implications of Distributed Energy Development in Blockchain-Based Institutional Environment," *Aptisi Trans. Technopreneursh.*, vol. 5, no. 2sp, pp. 209–220, 2023.
[7] Yang T, Jiang C, Li P. Video Analysis and System Construction of Basketball Game by Lightweight Deep Learning under the Internet of Things. Comput Intell Neurosci. 2022;2022.
[8] B. P. K. Bintoro, N. Lutfiani, and D. Julianingsih, "Analysis of the Effect of Service Quality on Company Reputation on Purchase Decisions for Professional Recruitment Services," *APTISI Trans. Manag.*, vol. 7, no. 1, pp. 35–41, 2023.
[9] Du Y, Zhao Q, Lu X. Semantic Extraction of Basketball Game Video Combining Domain Knowledge and In-Depth Features. Sci Program. 2021;2021.
[10] S. B. Goyal, E. P. Harahap, and N. A. Santoso, "Analysis of financial technology implementation on the quality of banking services in indonesia: Swot analysis," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 77–82, 2022.
[11] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):386–97.
[12] Q. Aini, I. Sembiring, A. Setiawan, I. Setiawan, and U. Rahardja, "Perceived Accuracy and User Behavior: Exploring the Impact of AI-Based Air Quality Detection Application (AIKU)," *Indones. J. Appl. Res.*, vol. 4, no. 3, pp. 209–218, 2023.
[13] Garza G. Mask R-CNN for Ship Detection & Segmentation. Towar Data Sci. 2019;
[14] Kulkarni KM, Shenoy S. Table tennis stroke recognition using two-dimensional human pose estimation. IEEE Comput Soc Conf Comput Vis Pattern Recognit Work. 2021;4571–9.
[15] U. Rahardja, Q. Aini, P. A. Sunarya, D. Manongga, and D. Julianingsih, "The Use of TensorFlow in Analyzing Air Quality Artificial Intelligence Predictions PM2. 5," *Aptisi Trans. Technopreneursh.*, vol. 4, no. 3, pp. 313–324, 2022.
[16] Hurault S, Ballester C, Haro G. Self-Supervised Small Soccer Player Detection and Tracking. MMSports 2020 - Proc 3rd Int Work Multimed Content Anal Sport. 2020;20:9–18.
[17] Wang T, Li T. Deep Learning-Based Football Player Detection in Videos. Comput Intell Neurosci. 2022;2022.
[18] Liu J. Convolutional Neural Network-Based Human Movement Recognition Algorithm in Sports Analysis. Front Psychol. 2021;12.
[19] Liu T, García-de-Alcaraz A, Wang H, Hu P, Chen Q. Impact of Scoring First on Match Outcome in the Chinese Football Super League. Front Psychol. 2021;12.
[20] Uchida I, Scott A, Shishido H, Kameda Y. Automated offside detection by spatio-temporal analysis of football videos. MMSports 2021 - Proc 4th Int Work Multimed Content Anal Sport co-located with ACM MM 2021. 2021;17–24.
[21] S. Mehta and L. Magdalena, "Education 4.0: Online Learning Management Using Education Smart Courses," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 70–76, 2022.
[22] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49.
[23] A. Gunawan and R. K. Hudiono, "Industrial Revolution 4.0's Information Technology's Impact on the Growth of MSMEs in the Manufacturing Industries Sector," *Int. Trans. Educ. Technol.*, vol. 1, no. 2, pp. 157–164, 2023.
[24] Liu H, Aderon C, Wagon N, Liu H, MacCall S, Gan Y. Deep Learning-based Automatic Player Identification and Logging in American Football Videos. 2022; Available from: http://arxiv.org/abs/2204.13809
[25] Wei CT, Weng SK. A court line extraction algorithm for badminton tournament videos with horizontal line projection learning. IET Image Process. 2023;17(10):2907–24.