International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2024, 8 (2), 76-80
DOI: 10.36287/ijmsit.8.2.5
Received: November 11, 2024; Accepted: December 9, 2024

A Prototype Study on YOLOv10-Based Bird Gesture Recognition

Rıdvan Yayla*

[1*]*Computer Engineering Department, Engineering Faculty Bilecik Şeyh Edebali University, Bilecik, Türkiye (ridvan.yayla@bilecik.edu.tr) (ORCID: 0000-0002-1105-9169)*

Abstract – Birds are one of the most abundant types of creatures on Earth. However, it is also known that many taxonomically diverse bird species exist in nature. The bird network has standard behavioural patterns such as flying, perching, feeding and walking. In this study, 2372 bird images are used for five standard bird gestures detection: flying, perching, swimming, eating, and walking with the Yolov10 algorithm from Caltech-UCSD Birds-200-2011 dataset. Firstly, the dataset is prepared for detection by classifying these gestures. Secondly, the bird gesture images are trained with Yolov10, thirdly the trained model is tested with bird motion short videos and finally, the evaluation results are shown with evaluation metrics. In this prototype study, it was observed that the obtained model had results with an accuracy higher than 70%. The study can be used to make sense of bird communication for future studies.

Keywords *– Bird gesture, target detection, classification, deep learning, Yolov10.*

Citation: Yayla, R., (2024). A Prototype Study on YOLOv10-Based Bird Gesture Recognition. International Journal of Multidisciplinary Studies and Innovative Technologies, 8(2): 76-80

I. INTRODUCTION

Object detection is one of the methods used in all intelligent systems in recent years. It is used for all intelligent systems from home management systems to target detection in military defense systems. In this way, the workload requiring manpower can be reduced. In recent years, a few effective object detection algorithms such as R-CNN, Fast R-CNN, Mask R-CNN and Yolo (You Only Look Once) have been developed for object detection. These methods are widely used for different purposes such as military defence, unmanned vehicles, agriculture, health and commerce.

On the other hand, a bird network is a widely complex structure in nature. According to ornithologists, 9500 to 11000 species of birds live in the world as a taxonomy [1]. In contrast, these animals have standard behaviour types such as flying, feeding and perching. In this scope, Yolov10 that is last version of the Yolo algorithm is used for bird gesture recognition in this study [2].

II. MATERIALS AND METHOD

A. Level Literature Review

In literature, there are various object detection studies with Yolo such as from traffic sign recognition to military target detection. Yolo algorithm is a flexible algorithm that can be used for a wide range of different objects. For bird studies, the Yolo algorithm has generally been used for purposes such as bird species recognition and bird flock detection for safe flight. Liang et al. developed the SMB-YOLOv5 model to detect birds near airports for safe flight in their study [3]. Datar et al. conducted a comparative study for bird detection using YOLOv2, YOLOv3 and Mask R-CNN algorithms [4]. Xie et al. used the yolov5 algorithm to detect the presence of bird nests that threaten transmission lines and transformer substations [5]. Ou et al. developed a system that detects birds and identifies bird species using Yolo and Custom Vision algorithms [6]. Zhao used the yolov4 algorithm to detect bird movements in 3 classes: staying, flying and swimming in a study she conducted in 2022 [7]. Bird gestures are improved by using Yolov10 for 5 movements: perching, swimming, opening wings-flying (ow-flying), eating, and walking in this study.

B. Yolo Algorithm

Yolo is the most common algorithm that uses CNN for real-time object tracking. Applications such as R-CNN, Fast R-CNN, and Faster R-CNN were popular applications used for real-time object tracking until Yolo was released in 2015 [8] [9] [10] [11].

Yolo predicts the coordinates and class of objects in the image by passing the image through a neural network one at a time. While making this definition, it divides the image into grids. There is no specific condition to determine the number of grids. It is sufficient to have it in N*N format. Each grid determines whether there is an object in it and if it detects an object, it estimates whether the center point is in its area. After the image passes through the neural network, a vector is produced as output. The working principle of YOLO is as follows:

- Yolo first divides the image into grids.
- It draws a frame (bounding box) around the objects
- in each region.
- It calculates the probability of finding an object in each region.
- It calculates a confidence score for each frame (bounding box). Confidence score predicts the

probability that an object is that object and is calculated as (1) [12].

$$Confidence = \Pr(obj) \times IoU \quad (1)$$

- In (1), Pr(obj) represents the probability of finding the object in the grid, and Intersection over Union (IoU) represents the intersection of the predicted box with the box where the object is located.

In YOLO, error functions are expressed with three basic error rates:

Confidence loss: It expresses how wrong it is to determine whether there is an object inside the grid. If there is an object in the image, it is calculated as (2) and if not, it is calculated as (3) [14].

$$\sum_{i=0}^{s^2}\sum_{j=0}^{B} 1_{ij}^{obj}\left(C_i - \hat{C}_i\right)^2 \quad (2)$$

$$\lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B} 1_{ij}^{noobj}\left(C_i - \hat{C}_i\right)^2 \quad (3)$$

S is the grid cell number, and B is the number of estimated bounding boxes in each grid cell. 1_{ij}^{obj} indicates whether the j bounding box of the i grid cell contains an object. λ_{noobj} is a weighting parameter used to compensate for the loss of confidence in places where objects are not present. C_i real confidence score i.e. the confidence score in the grid cell of the real object box and $\hat{C}_i$ is the confidence score estimated by the model.

Location loss: It expresses how wrong the predicted box is and it is calculated as (5) [13].

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} 1_{ij}^{obj}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} 1_{ij}^{obj}\left[\left(\sqrt{w_i} - \sqrt{\widehat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2\right] \quad (5)$$

λ_{coord} is a weighting parameter used to compensate for position loss. x_i and y_i indicate the x and y coordinates of the centre of the bounding box. $\hat{x}_i$ and $\hat{y}_i$ are the x and y coordinates of the centre of the bounding box predicted by the model. w_i , h_i represent the width and height of the bounding box, and $\widehat{w}_i$, $\hat{h}_i$ represent the width and height of the bounding box predicted by the model.

Classification loss: It expresses how wrong the predicted object is and is calculated as (4) [13].

$$\sum_{i=0}^{S^2} 1_i^{obj}\sum_{c\,\in\,classes}(p_i(c) - \hat{p}_i(c))^2 \quad (4)$$

$p_i(c)$ is a probability vector of real class labels, an array of 1 for the real object class and 0 for other classes. $\hat{p}_i(c)$ is the vector containing the class probabilities predicted by the model, that is, it represents the probability of the predicted class.

Finally, the total loss function is computed by summing three loss functions and it is shown as (6) [14]. The smaller the total loss function result, the higher the success of the algorithm.

Additionally, Mean Average Precision (mAP) is used to determine the accuracy of object identification by the Yolo algorithm [15]. mAP is a metric used to evaluate the accuracy of an object detection model. It measures the model's performance over a certain threshold value, averaged over all classes Average Precision (AP) is defined as the area under the precision and recall curve for a given class.

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} 1_{ij}^{obj}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} 1_{ij}^{obj}\left[\left(\sqrt{w_i} - \sqrt{\widehat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2\right] + \sum_{i=0}^{s^2}\sum_{j=0}^{B} 1_{ij}^{obj}\left(C_i - \hat{C}_i\right)^2 + \lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B} 1_{ij}^{noobj}\left(C_i - \hat{C}_i\right)^2 + \sum_{i=0}^{S^2} 1_i^{obj}\sum_{c\,\in\,classes}(p_i(c) - \hat{p}_i(c))^2 \quad (6)$$

IoU is the ratio of the intersection of the predicted boundary box and the true boundary box to the union area [12]. IoU is used to measure how accurate the model's detections are. mAP50 is a mAP calculation method where the IoU threshold is set to 0.5 (i.e. 50%). This means that if the intersection ratio of the predicted boundary box and the true boundary box is at least 50%, the prediction is considered accurate. The mAP50 value is the average of the AP values calculated for all classes and the IoU threshold of 0.

Moreover, mAP50-95 is another average AP value obtained by calculating the IoU threshold values from 0.5 to 0.95 with every 0.05 increments. The mAP50-95 metric measures how well the model performs not only at a 50% IoU threshold but also at higher IoU values. This provides a more challenging and detailed assessment. In the Yolo algorithm training process, these loss functions are computed by using Python programming.

III. System Design and Components

C. Dataset

2372 bird images has been used for bird gesture classification from the Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset [16]. 1161 train set (80%), 207 validation set (%10) and 207 test set (%10) images are used for the Yolo training process. Additionally, the dataset image number has been increased by using data augmentation with different views such as rotation, and mirroring. A few sample bird images with different gestures are shown in Figure 1.

D. Pre-process

Caltech dataset originally contains 11788 bird images with different species, views and poses. In this study, 2372 bird images have been used for five classes. Image data should be optimized for the training process. The bird images have been classified into 5 classes that are perching, opening wings-flying (ow-flying), swimming, eating, and walking in this study.

Each dataset image has been prepared for the training process. The images should be segmented for each defined class before the training process. The data set must be balanced and the epoch number must not be excessive to prevent over-learning during the training process.

250 images are selected for each movement class and each class has been enhanced by data augmentation from the Caltech dataset. These selected images are increased by using the Roboflow framework as balanced up to 2372 images.

Additionally, the training process is limited by 100 epochs to prevent over-fitting.

Fig. 1. Sample bird images in Caltech dataset [17][18]

Roboflow is a computer vision developer framework that focusses on improving data gathering, preprocessing, and model training procedures [18]. Data images are segmented for each class by using the Roboflow framework. Two sample segmented bird images from the Caltech dataset are shown in Figure 2.

Table 1. The dataset image numbers for each classes

Class	The number of selected images	The number of data augmentation images	The final numbers of images
eating	250	224	474
ow- flying	250	223	473
perching	250	226	476
swimming	250	220	470
walking	250	229	479
Total	1250	1122	2372

Fig. 2. Segmented sample bird images with RoboFlow framework[18][19]

Additionally, Roboflow provides to increase images for each class by using different techniques such as reflector, reversing, rotating right, left, above or below the image for data augmentation. In this way, the selected images from the Caltech dataset have been increased based on these five classes in this study. The numbers of images in the final dataset for each class are shown in Table 1.

E. Bird Gesture Detection

The study aims to determine bird gesture detection with the highest accuracy. In this scope, the dataset is prepared by using pre-processing techniques. The five-movement classes have been drawn by Roboflow and the final dataset has been prepared for the training process. The training has been made by NVIDIA L4 GPU and 52GB RAM, 22,5 GB GPU RAM and 78GB disk space on the Google Colab platform with 100 epochs and 8 batch sizes. At the end of the training process, a bird gesture detection trained model file has been obtained for 5 classes (perching, swimming, opening wings-flying (ow-flying), eating, and walking). In the test process, the detected bird gestures are shown in Figure 3 by using various bird video samples.

IV. RESULTS

The study is built based on the Yolov10 algorithm for detection bird gesture detection. mAP50 is one of the most commonly used metrics to evaluate the performance of an object detection model. It measures the average precision at IoU=0.5, taking into account both the accuracy of the objects detected by the model and the proportion of objects it misses. The accuracy metrics which are precision, and recall based on mAP scores are shown in Table 2.

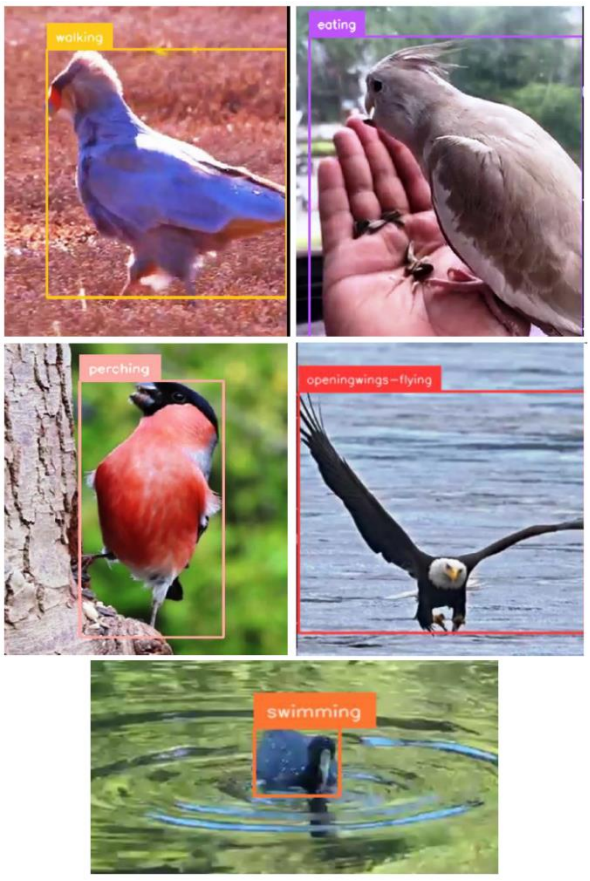

Fig. 3. Bird gesture detection samples for each class after training with the YOLOv10 model [20][21][22][23][24]

Table 2. Yolo Performance Metrics for each class

Class	Images	Instances	Performance Metrics			
			Precision	Recall	mAP50	mAP50-95
All	207	224	0.814	0.732	0.768	0.696
eating	207	39	0.833	0.640	0.756	0.688
ow- flying	207	43	0.943	0.860	0.913	0.882
perching	207	62	0.588	0.387	0.433	0.304
swimming	207	35	0.905	0.971	0.939	0.863
walking	207	45	0.800	0.800	0.799	0.743

Confusion Matrix is a table used to visualize and evaluate the prediction performance of a model. Confusion Matrix is used to better understand the classification or detection accuracy of a model [25]. Normalized Confusion Matrix is obtained by normalizing the value in each cell per class. Normalization is done by dividing the value in each cell by the total real count of that class. In this study, the normalized confusion matrix obtained as a result of training for five-bird gesture detection is shown in Figure 4.

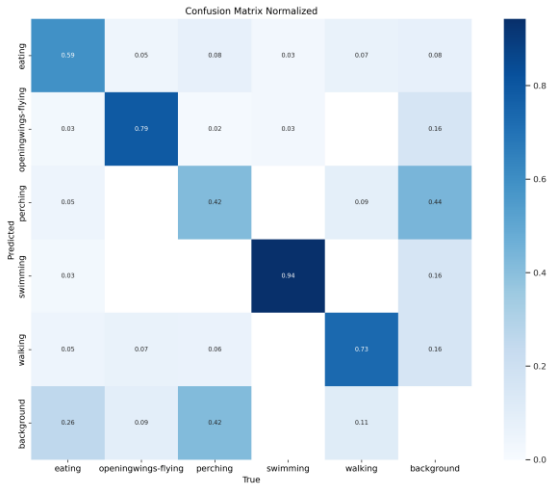

Fig. 4. Normalized confusion matrix for five classes

According to Table 2, the most challenging classes are perching and eating due to the similar pose of the bird. Because of the gesture similarity, the accuracy of these classes more less than the other classes. However, the perching position has been segmented with tree branches together to prevent gesture similarity. The best result was obtained from the flying position as it is a unique movement in birds. For the eating class, segmentation was made by taking into account whether the bird's beak is open or there is food in its beak. Because this segmentation is the most important feature that distinguishes the bird from the walking or perching class. In addition, since swimming birds such as ducks do not have visible feet on the water surface, better results can be obtained. Since the foot segmentation of the bird is particularly important in the walking position, the foot segmentation is highlighted together with the total segmentation of the bird in such images.

When mAP50 values are examined, it is seen that all classes except the perching position have an accuracy rate above 70%. In this study, a prototype study was conducted for bird gesture segmentation. The accuracy values can be increased with more balanced data and training to increase the accuracy rate of mAP50-95 values,

V. DISCUSSION

In this study, bird movements were classified into five different categories using the latest YOLO algorithm, YOLOv10, demonstrating its effectiveness in detecting and categorizing bird behaviours. Compared to previous studies, an increase in the number of movement classes was achieved, highlighting the potential of the proposed approach for capturing more granular behavioural distinctions. This advancement underscores the importance of leveraging state-of-the-art algorithms in ecological and behavioural research, as it allows for a more nuanced understanding and classification of animal behaviours.

However, despite the promising results, certain limitations were observed, primarily related to dataset size and balance. A balanced dataset is critical to avoid biases in classification performance across different movement categories. Future studies can enhance the dataset by including diverse and representative samples from various bird species and environments. This would not only improve the accuracy of movement detection but also enhance the generalizability of the model.

Additionally, integrating multimodal data such as bird sounds alongside visual data presents an exciting avenue for future research. Combining audio and visual modalities can enable a more comprehensive interpretation of bird communication, potentially leading to ground breaking insights into the interplay between movement and vocalization.

VI. CONCLUSION

This study successfully classified bird movements into five distinct categories using the advanced YOLOv10 algorithm, demonstrating the potential of deep learning models in ecological and behavioral research. By increasing the number of classes compared to previous studies, the model provides a more detailed understanding of bird behaviors. However, the findings also emphasize the importance of balanced datasets and adequate training to achieve optimal results. Future work integrating bird sounds and visual data could open new horizons in understanding bird communication, paving the way for more holistic and multimodal approaches in this field.

Authors' Contributions

The authors' contributions to the paper are equal.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics

REFERENCES

[1] T. Puiu, "How many birds are there in the world?" ZME Science: https://www.zmescience.com/feature-post/natural-sciences/animals/birds/how-many-birds-are-there-in-the-world/, 2023.

[2] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.
[3] H. Liang, X. Zhang, J. Kong, Z. Zhao, and K. Ma, "Smb-yolov5: A lightweight airport flying bird detection algorithm based on deep neural networks," IEEE Access, vol. 12, pp. 84 878–84 892, 2024.
[4] P. Datar, K. Jain, and B. Dhedhi, "Detection of birds in the wild using deep learning methods," in 2018 4th International Conference for Convergence in Technology (I2CT), 2018, pp. 1–4.
[5] M. Xie, X. Li, C. Zhao, and C. Xu, "Identification of bird nest based on yolov5 algorithm," in 2023 5th International Academic Exchange Conference on Science and Technology Innovation (IAECST), 2023, pp. 811–814.
[6] Y.-Q. Ou, C.-H. Lin, T.-C. Huang, and M.-F. Tsai, "Machine learning-based object recognition technology for bird identification system," in 2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). IEEE, 2020, pp. 1–2
[7] S. Zhao, "Bird movement recognition research based on yolov4 model," in 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 2022, pp. 441–444.*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
[8] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3520–3529.
[9] R. Girshick, "Fast r-cnn," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1137– 1149, 2016.
[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[12] P. Shivaprasad, "A comprehensive guide to object detection using yolo framework," https://towardsdatascience.com/a-comprehensive-guide-to-object-detection-using-yolo-framework-24f8e2e5c6ab, jan 2019.
[13] E. Yildirim, U. G. Sefercik, and T. Kavzoglu, "Automated identification of vehicles in very high-resolution uav orthomosaics using yolov7 deep learning model." Turkish J. Electr. Eng. Comput. Sci., vol. 32, no. 1, pp. 144–165, 2024.
[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 779–788. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.91
[15] Y. Indulkar, "Alleviation of covid by means of social distancing face mask detection using yolo v4," in 2021 International Conference on Communication information and Computing Technology (ICCICT), 2021, pp. 1–8.
[16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "Caltech-ucsd birds-200-2011 (cub-200-2011)," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
[17] Wang, F., Zhou, H., Li, S., Lei, J., & Zhang, J. (2020). Convolutional Attention Network with Maximizing Mutual Information for Fine-Grained Image Classification. Symmetry, 12(9), 1511.
[18] B. Dwyer, J. Nelson, T. Hansen, and et al., "Roboflow (version 1.0)," https://roboflow.com, 2024, computer vision software.
[19] Shandilya, S. K., Srivastav, A., Yemets, K., Datta, A., & Nagar, A. K. (2023). YOLO-based segmented dataset for drone vs. bird detection for deep and machine learning algorithms. Data in Brief, 50, 109355.
[20] HAWIStudios. (2024, Jan 17). Caracara bird walking in field. [Short video]. https://youtube.com/shorts/fGuZgvVb3uc?si=AYq2f8v93TthqseH
[21] DiscoverAnimalAll. (2024, Jul 21). Bird Perched On A Tree While Eating. [Short video]. https://youtube.com/shorts/Qk53nb8kEQM?si=f0EUK-u_73BTRf8s
[22] devang_jani. (2023, Nov 27). Bird eating. [Reel]. Instagram. https://www.instagram.com/devang_jani/reel/C0I5L3wpkfH/
[23] MarkSmithphotography. (2023, June 30).Must SEE!! Extreme close up off a Bald Eagle snatching a fish from a whirlpool.[Short video]. https://youtube.com/shorts/saoQHEJCkBM?si=OHzeoK6BLCsiRTKX
[24] @recep.kaplan. (2023, Nov 8). Ankara Gölbaşı Mogan Gölü'nde Ördekler. [Short video]. https://youtube.com/shorts/aGSCvYjjt9I?si=mWdb6y3j9E1tpVUC
[25] N. Herbaz, H. El Idrissi, and A. Badri, "Deep learning empowered hand gesture recognition: using yolo techniques," in 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), 2023, pp. 1–7.