IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CNN-GRU Based Hybrid Architecture for Video Classification of Birds

[1]Sudharsan K, [2]Gowtham M, [3]Vanitha Ravi
[1]SQL Developer, [2]SDET, [3]Business Analyst
[1]Computer Science and Engineering,
[1]Coimbatore Institute of Technology, Coimbatore, India

Abstract: Bird classification plays a critical role in ornithology and wildlife conservation. With the rise of video data, the demand for efficient and accurate video-based bird classification methods has increased. This research introduces a hybrid architecture combining Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) for bird video classification. The CNN extracts spatial features from video frames, while the GRU models the temporal dependencies. The combination of global and local feature extraction enables the model to capture both fine-grained and coarse-grained bird behaviors, resulting in improved classification performance.

INTRODUCTION

Bird classification is a crucial task in ornithology and wildlife conservation. The increasing availability of video data has spurred demand for accurate video-based classification methods. Deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated significant success in tasks like image classification and video analysis.

The proposed architecture combines a CNN-based feature extractor with a GRU-based temporal modeling module. The CNN processes frames from bird videos, extracting high-level spatial features, which are passed to the GRU to model temporal dependencies. This hybrid approach captures both fine details like bird color patterns and overall motion dynamics, improving bird classification accuracy.

This model adapts well to different species and environments and requires smaller labeled datasets, making it highly suitable for various bird classification tasks.

NEED OF THE STUDY.

Traditional methods of bird identification often rely on static images or audio recordings, leaving out important dynamic information present in video data. With the increasing use of video footage for monitoring wildlife, there is a pressing need for effective and accurate methods for bird species identification through video.

Video-based bird classification is particularly important in scenarios where birds pose risks to human activities, such as aviation safety. Birds flying at certain altitudes near airports can pose significant hazards to aircraft during landing and takeoff. Therefore, developing a system that not only classifies birds but also determines their flight height is essential for improving aviation safety protocols.

This study aims to bridge the gap by introducing a CNN-GRU hybrid architecture that can analyze both spatial and temporal aspects of bird behavior from video data. By developing a robust bird detection and classification system, the study provides a real-time solution to assess bird flight safety in environments such as airports, where timely and accurate classification is critical.

OBJECTIVE:

The objective of this project is to develop a bird detection and video classification system that can accurately detect birds from video footage and classify them based on their species using a database of bird species with their average flight height. The key goals include:

Bird Detection: To Develop an algorithm to analyze video footage and detect birds based on visual characteristics such as size, shape, and movement patterns.

Video Classification: To Build a model to classify detected birds into their respective species using a large dataset of labeled bird videos.

Flight Height Comparison: To Create a mechanism to compare the detected bird's flight height with the average flight height of its classified species.

Safety Determination: To Assess if the detected bird's altitude is safe for airplane operations based on the comparison.

Integration: To Combine the modules into a cohesive system that processes video footage in real-time.

3.1Population and Sample

The population of this study consists of video footage of bird species taken from the VB100 dataset, a comprehensive dataset designed for fine-grained classification and deep learning experiments. The dataset contains a total of 1,416 Video clips of 100 different bird species, making it suitable for video-based bird classification tasks. Each species is represented by an average of 14 video clips, with a median video length of 32 seconds. The dataset introduces several challenges such as variations in scale, pose, background, and camera movement, which are common in real-world bird monitoring scenarios.For the training and testing phases of the model, the videos are divided into two subsets: 70% for training and 30% for testing. This division ensures that a sufficient amount of data is used to train the model, while the remainder is

reserved for evaluating the model's performance. The dataset is specifically chosen to challenge the model's ability to generalize across bird species and handle variations in environmental factors such as camera motion and frame rate, which range from 25 fps to 100 fp

3.2 Data and Sources of Data

The data for this study is derived from the VB100 dataset, which is a publicly available dataset curated by expert bird watchers for fine-grained classification tasks in computer vision. The dataset contains a total of 1,416 video clips from 100 bird species. These video clips are accompanied by metadata such as bird taxonomy, geographical distribution, and other relevant characteristics. Each bird species has an average of 14 video clips, with a median video length of 32 seconds, making it an ideal dataset for deep learning and video classification experiments. The VB100 dataset presents challenges such as variations in frame rates (ranging from 25 fps to 100 fps), camera movement, and varying environmental conditions. Approximately 69% of the videos were captured at 30 frames per second (fps), while the remainder were captured at 25, 60, or 100 fps. These variations make the dataset suitable for testing the robustness of bird classification models under real-world conditions.The data is split into training and testing sets, with 70% of the video clips used for training the bird detection and classification models, and the remaining 30% reserved for testing and evaluation. This split ensures that the models are trained on a diverse set of bird species and behaviors, while still being tested on unseen data to assess their generalization capabilities.

3.3 Theoretical framework

The proposed bird classification model utilizes a hybrid deep learning architecture, combining Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) to classify bird species based on video data. This hybrid approach leverages both spatial and temporal features inherent in video data. Below are the key components of the theoretical framework:

1. **Convolutional Neural Networks (CNNs):** CNNs are widely used in image processing and computer vision tasks. In this study, the CNN serves as the feature extractor, processing individual frames from the bird videos to capture spatial features such as color patterns, shapes, and textures. By extracting high-level spatial information, the CNN helps detect key visual characteristics of birds, such as body parts, wings, and movement trajectories.
2. **Gated Recurrent Units (GRUs)**: GRUs are a type of Recurrent Neural Network (RNN) designed to handle sequential data. Unlike Long Short-Term Memory (LSTM) networks, GRUs are computationally simpler and faster, making them more suitable for handling the temporal dependencies present in video data. GRUs model the temporal dynamics of the extracted features from CNN, capturing motion patterns, bird behavior sequences, and flight trajectories over time. This enables the model to recognize complex temporal relationships between frames.
3. **Hybrid CNN-GRU Architecture:** The hybrid model is built by connecting the CNN feature extractor with the GRU temporal modeling unit. This architecture allows the system to process both spatial (frame-level) and temporal (sequence-level) information from bird videos, leading to a more comprehensive understanding of bird species and their behaviors. The dual focus on spatial and temporal features improves classification accuracy, especially when handling diverse bird species with varying motion patterns.
4. **Feature Selection and Processing:** The CNN extracts spatial features from each video frame, which are then passed through the GRU to model the temporal sequence of the frames. ResNet50, a lightweight CNN architecture, is used in the classification module, while MobileNet is employed for bird detection. These models are chosen due to their ability to perform well on large-scale datasets with limited computational resources, ensuring that the hybrid architecture can be deployed efficiently in real-time systems.
5. **Flight Height Determination**: In addition to classification, the system incorporates a comparison mechanism for flight height determination. The detected bird's flight height is compared with the species' average flight height from a pre-existing bird database. This feature is particularly useful for determining whether a bird poses a risk to airplane operations, thereby integrating wildlife monitoring with aviation safety protocols.

The combination of these deep learning techniques forms the foundation for accurate and efficient video-based bird classification, allowing the model to handle complex video data while minimizing training time and computational resources.

3.4Statistical tools and econometric models

To evaluate the performance of the CNN-GRU hybrid architecture for video-based bird classification, various statistical and econometric methods are employed:

3.4.1 Descriptive Statistics: Descriptive statistics such as mean, standard deviation, and variance are used to summarize the key attributes of the VB100 dataset, such as the distribution of video lengths and frame rates. The Jarque-Bera test is applied to check the normality of the data.

3.4.2 Model Construction

1. **Data Collection:** The VB100 dataset consists of 1,416 video clips representing 100 different bird species. To ensure robust model training and evaluation, the dataset is split into a 70:30 ratio, allocating 70% of the clips for training and 30% for testing. This allows the model to learn from a substantial portion of the data while retaining a separate set for validation of its performance.
2. **Data Preprocessing:** In the **data preprocessing** stage, the video frames are resized and **cropped** to **224x224 pixels** to standardize the input size for the model. This ensures uniformity across the dataset, which is crucial for effective training. Each video is organized into batches containing a fixed number of frames, and any shorter videos are **padded** to maintain consistent input dimensions. This step prepares the data for efficient processing by the neural network.
3. **Feature Extraction:** For **feature extraction**, two pre-trained models are employed:

- MobileNet is utilized for bird detection, focusing on identifying birds within the video frames. It is lightweight and optimized for mobile devices, making it suitable for real-time applications.
- ResNet50 is used for bird species classification, as it effectively captures fine-grained details of bird species from the detected images. This model is known for its high accuracy and efficiency, enabling it to classify a large number of species based on the extracted features.
- The architecture of the model combines Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs):

- The CNN is responsible for extracting spatial features from the video frames, allowing the model to learn characteristics such as shape, color, and texture of the birds.
- The GRU is employed to capture temporal dependencies across frames, effectively modeling the movement and behavior of birds over time. This temporal aspect is crucial for understanding bird dynamics and enhances the model's ability to classify species based on their actions within the videos.

Econometrics

1. The econometric analysis is used to compare the relative performance of the CNN architectures.

table :1

	loss	**accuracy**	**precision**	**recall**
VGG16	4.205457687	77.08333135	0.7708333135	0.7708333135
VGG19	5.550269604	79.16666865	0.7916666865	0.7916666865
Resnet50	1.563070059	84.0277791	0.840277791	0.840277791
Densenet	4.547659397	36.80555522	0.3834586442	0.3834586442
Inception	11.13249779	24.30555522	0.2517985702	0.2430555522
EfficientNet	1.835297465	53.4722209	0.5757575631	0.527777791

2. ResNet50 demonstrates superior performance in the bird classification task, as indicated by its lower loss value, which suggests effective generalization to unseen data. In contrast, Inception Net and DenseNet show poor performance across all metrics, making them less suitable for this specific classification challenge. Both VGG16 and VGG19 also performed well, with VGG19 achieving a higher accuracy of 79.17% compared to VGG16's 77.08%. However, their higher loss values relative to ResNet50 imply that, while capable of making accurate predictions, these models may be more prone to overfitting or require further tuning for improved generalization.

3. **Model Selection:** Based on the econometric comparison, ResNet50 is identified as the optimal model for bird video classification due to its superior performance in all key metrics. ResNet50, is the best in terms of accuracy, is chosen as the final model for deployment due to its balance of accuracy(84.02%) and lightweight architecture, making it more suitable for real-time applications on devices with limited computational resources.

4. **Training Time and Model Efficiency**: Training time for different CNN architectures (EfficientNetV2, VGG16, DenseNet) was measured and compared. ResNet50 was selected based on its optimal balance between accuracy and speed. The training process, conducted using Google Colab's free GPU, averaged 10-15 minutes per model. DenseNet was the fastest to train; however, ResNet50 outperformed it in terms of accuracy.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables:

The bird detection module demonstrated effective performance in identifying birds within the video frames of the VB100 dataset. Through rigorous testing and evaluation, ResNet-50 emerged as the most suitable architecture for the classification task. Initially, several convolutional neural network architectures were explored, including VGG16, VGG19, DenseNet, and EfficientNet. Each model was evaluated based on its ability to accurately classify the different bird species present in the dataset. Among these architectures, ResNet-50 exhibited the highest accuracy, achieving an impressive accuracy rate of around 84%. This result highlights ResNet-50 has capability to balance model complexity and accuracy, making it an effective choice for the classification of bird species from video data.While DenseNet showed promise due to its rapid training times, it did not match the classification accuracy of Resnet . The superior performance of this can be attributed to its efficient scaling and the way it handles multi-resolution features, allowing it to capture intricate details necessary for distinguishing between species. Overall, the findings suggest that Resnet is a robust model for bird species classification in video frames, outperforming other established architectures both in accuracy and overall effectiveness for this specific application. This work not only contributes to the understanding of avian biodiversity but also sets a foundation for future research in wildlife monitoring and conservation through advanced machine learning techniques.

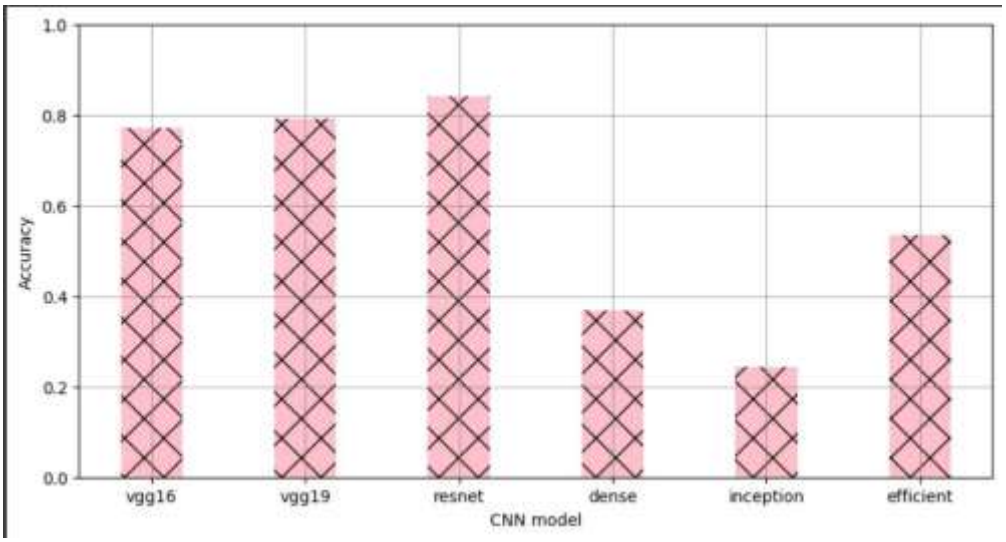

figure 1

Acknowledgment: We express my sincere gratitude to Dr. S.P. Abirami, M.E, Ph.D., Assistant Professor in the department of computer science and engineering at Coimbatore institute of technology for her invaluable guidance and support throughout this research. Her expertise and insights have significantly contributed to the development of this work. I also thank the International Journal of Novel Research and Development for providing a platform to publish my findings.

References:

1. Atanbori, J., Duan, W., Shaw, E., Appiah, K., & Dickinson, P. (2018). Classification of bird species from video using appearance and motion features. *Ecological Informatics*, *48*, 12-23.
2. Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). Ieee.
3. Rozental, A., & Fleischer, D. (2018). Amobee at semeval-2018 task 1: Gru neural network with a cnn attention mechanism for sentiment classification. *arXiv preprint arXiv:1804.04380*.
4. Dua, N., Singh, S. N., & Semwal, V. B. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. *Computing*, *103*(7), 1461-1478.