A Multi-Stage Framework for Detection and Decoding of Wildlife Ring Identifiers in Real-World Conditions

Tom Bourbon[1], Thomas Blanchon[2], Jocelyn Champagnon[2], and Cédric Pradalier[1]

[1] GeorgiaTech Europe – CNRS IRL 2958, Metz, France
[2] Tour du Valat, Research Institute for the Conservation of Mediterranean Wetlands, Arles, France

Abstract. Recent advances in automated wildlife monitoring have greatly increased the availability of ecological image data, creating a strong need for scalable analysis methods. In this work, we address the automated identification of Eurasian spoonbills (Platalea leucorodia) using camera trap imagery collected in long-term monitoring programs in the Camargue (Southern France). Each bird is equipped with a unique four-character PVC ring, enabling individual identification.
We propose a deep learning-based pipeline that automatically detects and reads alphanumeric ring codes directly from images. This automated approach is intended to accelerate the image-by-image review process. We evaluate our method against human annotations and demonstrate its effectiveness for large-scale individual identification in wild bird populations.

1 Introduction

Methods of data collection in biology have undergone a rapid evolution in recent years, with a clear shift toward automation. Technologies such as drones for aerial imaging [1] and strategically deployed camera traps [2] now enable the acquisition of massive volumes of ecological data. While these advances significantly improve monitoring capabilities, they also generate large volumes of data that are challenging to process and analyze efficiently [3]. Manual analysis is increasingly impractical due to its repetitive and time-consuming nature, creating a strong need for automated or semi-automated solutions.

Artificial intelligence, and particularly machine learning, has emerged as a key tool to address this challenge. In biology, these methods are widely used to automate tasks such as species detection, counting, and classification. Applications include the enumeration and identification of microorganisms [4] as well as macro-organisms [5]. However, beyond species-level analysis, identifying individuals within populations remains a critical objective for ecological studies, particularly for estimating survival rates, tracking movement patterns, and understanding population connectivity [6].

In this work, we focus on the automated identification of individual Eurasian Spoonbills (*Platalea leucorodia*, hereafter spoonbill) within breeding colonies. Since 2008, a long-term monitoring program in the Camargue (Southern France) has relied on rings engraved with unique alphanumeric codes to identify individuals. Currently, images collected through camera traps are analyzed manually by experts and through participatory science initiatives, such as the *Where is Spoony* project on the Zooniverse platform. This process is labor-intensive and limits scalability.

This study introduces an automated deep learning pipeline for the detection and recognition of alphanumeric codes on bird-mounted tags directly from images, leveraging existing marking systems without additional animal intervention. We demonstrate its effectiveness through a comparative evaluation against results obtained via citizen science.

2 State of the Art

The integration of machine learning into biological data analysis has led to significant advances in automated species identification and counting. Applications range from microorganism detection [4] to the classification of macro-organisms in complex environments [5]. These approaches have demonstrated strong performance in handling large-scale datasets generated by modern acquisition systems, particularly through the use of deep convolutional architectures and large annotated datasets.

At the individual level, several image-based identification methods have been proposed. Many rely on natural, individual-specific visual features, such as unique patterns on whale skin [7] or distinctive feather markings in birds [8]. While effective, these methods are limited to species with sufficiently discriminative and stable natural patterns, and may suffer from sensitivity to pose, occlusion, and environmental variability. Furthermore, their performance can degrade in long-term monitoring scenarios where appearance may evolve over time.

Alternative approaches involve artificial tagging systems. Traditional methods include attaching rings or markers with unique codes to individuals, which can then be identified visually [9]. Recent work has explored automated detection of such tags in images, enabling researchers to filter large datasets and focus on relevant observations [10]. However, these methods typically focus on detection alone and still require manual transcription of the alphanumeric codes, which remains a major bottleneck for large-scale studies. This limitation significantly constrains the scalability of such approaches in long-term ecological monitoring programs.

From a computer vision perspective, the problem of reading alphanumeric tags is closely related to scene text detection and recognition. Significant progress has been made in this area, with methods combining object detection and sequence recognition to extract text from natural images. Recent architectures based on attention mechanisms and transformer models have further improved recognition accuracy in complex scenes. However, directly applying these ap-

proaches to ecological data is challenging due to the small size of the targets, low resolution, motion blur, and highly variable acquisition conditions encountered in field deployments.

Other sensing modalities have also been investigated, such as acoustic-based species identification using machine learning [11]. While effective at the species level, these approaches generally do not provide individual-level discrimination and are therefore not suitable for tracking tagged organisms.

Despite these advances, the joint problem of automatic detection and reliable reading of alphanumeric identification tags in unconstrained ecological settings remains relatively underexplored. Addressing this gap is essential for reducing manual processing effort and enabling scalable, long-term monitoring of individuals in natural environments. In particular, approaches that jointly address detection, geometric normalization, and sequence recognition are likely to play a key role in advancing automated ecological monitoring systems.

3 Methodology

3.1 Data collection

This section describes the data acquisition and annotation processes used to construct the dataset. Particular attention is given to the challenges inherent to real-world outdoor imaging conditions and to the strategies employed to ensure high-quality ground-truth labels.

Image collection. The dataset consists of high-resolution images (5760×4320 pixels) acquired using fixed camera traps under unconstrained outdoor conditions. The images contain small regions of interest corresponding to structured alphanumeric identifiers. The acquisition setup introduces substantial variability, including scale changes, motion blur, partial occlusions, and significant illumination variations. These factors contribute to the complexity of the visual recognition task. An example of a typical camera trap image is provided in Fig. 1, while Fig. 2 illustrates the diversity of challenging scenarios encountered in practice through a representative set of difficult conditions.

Label collection. Ground-truth annotations are collected via the citizen science platform *Zooniverse*, where multiple participants independently review each image and identify visible tags along with their associated alphanumeric codes. After completing each annotation, contributors provide a self-assessed confidence score. To ensure annotation reliability, only labels with a confidence score greater than or equal to 4 out of 5 are retained. This threshold was selected as it yields a level of agreement comparable to expert annotations while preserving a sufficiently large dataset. This filtering process mitigates the impact of noisy or uncertain labels and enhances the overall quality of the ground-truth annotations used for evaluation.

Fig. 1. Example of a typical image acquired by the camera trap system under natural conditions.

3.2 Proposed Approach

We propose a deep learning pipeline organized into three stages: detection, cropping, and reading. First, an object detection model localizes candidate tag regions in high-resolution images. The detected regions are then cropped and passed to a recognition module, which performs alphanumeric sequence decoding.

This staged design improves robustness by separating localization from recognition, ensuring that only relevant regions are processed during the reading step. As a result, the system reduces erroneous predictions arising from background clutter and focuses the recognition process on valid tag areas.

3.3 Stage 1 (Detection)

The first stage aims at coarse localization of regions containing alphanumeric codes and is formulated as an object detection problem. We adopt a YOLOv8-based convolutional neural network, chosen for its strong performance and efficiency in general object detection tasks.

The model is trained on 1,298 manually annotated images, where each annotation corresponds to a bounding box indicating the position of a tag within the scene. The main challenge lies in the extreme scale discrepancy between the objects of interest (approximately 32×96 pixels) and the full image resolution, as well as the high variability of visual conditions.

To address these issues, we use the SAHI framework [12], which enables detection on high-resolution images by dividing them into overlapping 640×640 patches extracted from the original full-resolution images. The YOLO detector is trained on such 640×640 windows, ensuring consistency between training and inference. During inference, predictions are also performed on these small

Fig. 2. Examples of challenging acquisition conditions: clear reading (top left), blur (top right), night-time capture (bottom left), and occlusion (bottom right).

patches, improving sensitivity to small objects while maintaining global coverage of the scene. The resulting detections from overlapping regions are then merged to produce final outputs at the original image scale.

3.4 Stage 2 (Cropping)

The second stage refines the localization obtained in the previous step by predicting oriented bounding boxes (OBB). Unlike axis-aligned detections, OBBs enable a tighter spatial fit around the target region while preserving its geometric orientation. This serves two main purposes: (i) removing irrelevant surrounding content to produce clean crops of the tag region, and (ii) estimating the in-plane rotation of the object, which facilitates the subsequent recognition stage.

This stage is implemented independently from the first one, as oriented bounding boxes are not supported by the SAHI framework used for large-scale patch-based inference. We therefore directly formulate this step as an oriented detection problem and employ a YOLOv8-OBB architecture to regress rotated bounding boxes.

3.5 Stage 3 (Reading)

Reading is performed using a single model operating on cropped images of previously detected oriented rings. The angle estimated in the detection stage is used to normalize ring orientation, effectively cancelling rotation so that all inputs to

the reading stage are aligned at either 0° or 180°. This orientation normalization is crucial for accurate alphanumeric classification, as characters such as "A" become fundamentally different when rotated (e.g., an A at 75° versus at 195°).

In this third-stage model, a YOLOv8 architecture is used to detect and classify the different codes present on each isolated ring.

Ring code letters are not uniformly distributed. They follow a structured generation process in which codes are assigned alphabetically and sequentially over time. Moreover, to coordinate international ringing programs and avoid duplicates, each national program typically uses a restricted subset of codes, often defined by a specific initial letter and ring color (e.g., in France for Spoonbills, codes start with letters A and F and use white font rings). As a consequence, a strong bias exists in the dataset, with most detected rings starting with the letter "A", leading to improved performance for this class at the expense of less frequent characters.

To mitigate this imbalance, synthetic ring data are generated in 3D using the OpenGL library. This approach allows the creation of approximately 200 rings per second without manual annotation, as labels are automatically generated during rendering. It also enables the construction of a uniform character distribution. Examples of the resulting synthetic objects after rendering and degradation are illustrated in Fig. 3.

To better match real-world conditions, synthetic images are further augmented by applying transformations such as down scaling, motion blur, coarse dropout (to simulate occlusions), and color jittering. Additionally, realistic shading and material properties of the cylindrical ring are directly handled in OpenGL.

The model is trained to recognize characters in both upright and inverted orientations, enabling rotation-invariant character recognition. This property is critical for inferring the global reading direction of the ring and ensuring correct sequence reconstruction. In particular, orientation is leveraged as an additional cue for disambiguating the reading order. When at least two characters are classified as inverted, the system infers that the ring is globally flipped, and consequently reverses the reading direction prior to reconstructing the full alphanumeric code. Characters invariant under horizontal inversion (e.g., H, X, N) are not considered for the estimation of the reading direction, as they do not provide reliable orientation cues.

A complete pipeline connecting the three stages is implemented, complemented by two post-processing filters designed to reduce false positives in the final system. The first filter verifies whether the predicted code tag belongs to a known list of valid tags, when available, while the second refines predictions based on the confidence score. In this second filtering stage, each individual character prediction is considered successfully read only if its classification confidence exceeds a threshold of 0.72. This threshold is applied at the per-letter level and was selected manually in an empirical manner through validation experiments, with the objective of reducing recognition errors caused by low-confidence predictions. By discarding uncertain character-level outputs, this step improves the reliability of the final reconstructed tag sequence.

This stage is trained on 20,150 images, including 2,156 real samples (10.7%) and 17,994 synthetic samples (89.3%).

Fig. 3. Examples of synthetic ring objects after rendering and degradation under realistic conditions, including blur, illumination variations, noise, and occlusions.

4 Results

4.1 Evaluation methodology

The proposed pipeline was evaluated on a dataset of 1,544 images collected from multiple cameras deployed at different locations across several colonies over a continuous period of 11 days. This setup introduces variability in viewpoints and acquisition conditions, while still capturing natural variations in illumination, occlusion, and object appearance over time.

On those 1544 images, a total of 1,759 rings were annotated by human via Zooniverse citizen science platform, considering only annotations with a confidence score of at least 4 out of 5. These annotations serve as the reference ground truth for the evaluation.

4.2 Detection Stage

The detection stage was evaluated on the same dataset of 1,544 images. In total, the detector identified 1,373 candidate rings out of the 1,759 annotated instances.

Importantly, no false positives were observed at this stage, indicating a high precision of the detection module. However, 386 annotated rings were not detected, corresponding to missed detections. This results in a recall of approximately 78.1%.

The absence of false positives suggests that the detector operates conservatively, favoring precision over recall. While this behavior is beneficial for limiting error propagation in subsequent stages, it also highlights that a non-negligible portion of rings remains undetected.

Qualitative analysis shows that most missed detections occur in challenging conditions, including severe occlusions, motion blur, or low-contrast situations where the visual appearance of the rings is significantly degraded. These limitations point to potential areas for improvement, particularly in enhancing the robustness of the detector under adverse imaging conditions.

4.3 Reading

Among these detections, 836 were correctly and fully read, corresponding to a precision of 61% with respect to the total number of detected rings. This result highlights the difficulty of the task, particularly due to partial occlusions, motion blur, and challenging lighting conditions that affect both detection and recognition stages.

To improve robustness, two successive filtering stages were applied to the raw detections. These filters were designed to eliminate spurious detections while preserving as many true positives as possible. After applying both filters described in Section 3.5, only a single false positive remained across the entire dataset of 1,544 images. This is comparable in magnitude to human reading performance, typically reported at around 1 error per 1000 reads, and demonstrates the effectiveness of the proposed filtering strategy.

Overall, the results indicate that the proposed pipeline achieves a strong balance between detection accuracy and robustness, with a very low false positive rate and a substantial proportion of correctly decoded rings among the detected candidates.

5 Discussion

We compare the performance of the proposed deep learning pipeline with annotations obtained through the Zooniverse citizen science platform. The model correctly and fully reads 48% of the images.

These results suggest that the system can be effectively integrated into manual annotation workflows. By providing model predictions directly to human annotators, the labeling process becomes less tedious and more efficient, effectively transforming the tool into a semi-automatic annotation system. In particular, uncertain predictions can be flagged for human validation. We implemented such an interface, which reduces by a factor of four the number of images that must be manually processed.

The proposed pipeline can also be deployed in a near real-time setting, reaching up to 35 frames per minute under the current configuration.

The approach is generalizable to other species marked with rings or tags. Most of the required adaptation lies in the synthetic data generation process, particularly for the reading stage, where character sets and visual appearance must be adjusted.

The semi-automatic design, which requests human confirmation for uncertain predictions, significantly reduces false positives. This is a critical property, as the resulting annotations are directly used in biological research databases, where reliability is essential.

Despite these advantages, the current system has several limitations. It remains sensitive to variations in font style, character sets (e.g., inclusion of digits), ring color, material properties, and image quality.

A promising direction for future work is the development of a unified model capable of reading multiple types of animal marking systems, including rings, tags, and other identification devices across species.

6 Conclusion

In this work, we present a complete three-stage pipeline for the detection and reading of bird ring codes in challenging real-world conditions. The proposed system combines oriented detection, geometrically consistent cropping, and a dedicated recognition model trained on a mix of real and synthetic data.

Experimental results show that the detection stage achieves high precision while maintaining a recall of 78.1%. The full pipeline is able to correctly and fully read 49% of the evaluated images, demonstrating its effectiveness under difficult acquisition conditions.

Beyond performance metrics, the system is designed to operate as a semi-automatic annotation tool, significantly reducing the manual effort required for large-scale ecological data processing. Its low false positive rate (1 in 1000) makes it particularly suitable for integration into scientific workflows where annotation reliability is critical. Overall, this work demonstrates the feasibility of an end-to-end deep learning approach for automated reading of wildlife identification rings, and opens the way toward more general multi-species and multi-marker recognition systems.

Acknowledgements

References

1. Jeffrey K Gillan, Guillermo E Ponce-Campos, Tyson L Swetnam, Alessandra Gorlier, Philip Heilman, and Mitchel P McClaran. Innovations to expand drone data collection and analysis for rangeland monitoring. *Ecosphere*, 12(7):e03649, 2021.
2. Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers. Use of camera traps for wildlife studies: a review. *Biotechnologie, Agronomie, Société et Environnement*, 18(3), 2014.
3. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of business research*, 70:263–286, 2017.
4. Aishwarya Venkataramanan, Martin Laviale, Cécile Figus, Philippe Usseglio-Polatera, and Cédric Pradalier. Tackling inter-class similarity and intra-class variance for microscopic image-based classification. In *International conference on computer vision systems*, pages 93–103. Springer, 2021.
5. Alexandre Delplanque, Samuel Foucher, Philippe Lejeune, Julie Linchant, and Jérôme Théau. Multispecies detection and identification of african mammals in aerial imagery using convolutional neural networks. *Remote Sensing in Ecology and Conservation*, 8(2):166–179, 2022.
6. Jean-Dominique Lebreton, Kenneth P Burnham, Jean Clobert, and David R Anderson. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological monographs*, 62(1):67–118, 1992.

7. Luis Fernando De Mingo Lopez, Clemencio Morales Lucas, Nuria Gómez Blas, and Krassimira Ivanova. Pattern recognition with convolutional neural networks: humpback whale tails. 2019.
8. André C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9):1072–1085, 2020.
9. Gustavo Alarcón-Nieto, Jacob M Graving, James A Klarevas-Irby, Adriana A Maldonado-Chaparro, Inge Mueller, and Damien R Farine. An automated barcode tracking system for behavioural studies in birds. *Methods in Ecology and Evolution*, 9(6):1536–1547, 2018.
10. Andrea Santangeli, Yuxuan Chen, Mark Boorman, Sofía Sales Ligero, and Guillermo Albert García. Semi-automated detection of tagged animals from camera trap images using artificial intelligence. *Ibis*, 164(4):1123–1131, 2022.
11. Fan Yang, Ying Jiang, and Yue Xu. Design of bird sound recognition model based on lightweight. *Ieee Access*, 10:85189–85198, 2022.
12. Naufal Najiv Zhorif, Rahmat Kenzie Anandyto, Albrizy Ullaya Rusyadi, and Edy Irwansyah. Implementation of slicing aided hyper inference (sahi) in yolov8 to counting oil palm trees using high-resolution aerial imagery data. *Int. J. Adv. Comput. Sci. Appl*, 15(7):1–6, 2024.