Deep learning-based methods for individual recognition in small birds

André C. Ferreira[a,b,c*], Liliana R. Silva[b,d], Francesco Renna[e], Hanja B. Brandl[c,f,g], Julien P. Renoult[a], Damien R. Farine[c,f,g], Rita Covas[b,h] and Claire Doutrelant[a,h]

[a]Centre d'Ecologie Fonctionnelle et Evolutive, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul-Valery Montpellier 3, Montpellier , France

[b]CIBIO-InBio, Research Centre in Biodiversity and Genetic Resources, Campus Agrário de Vairão, Vairão, Portugal

[c]Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany

[d]Université Paris-Saclay, CNRS, Institut des Neurosciences Paris-Saclay, 91190, Gif-sur-Yvette, France

[e]Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal

[f]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany

[g]Department of Biology, University of Konstanz, Germany

[h]FitzPatrick Institute of African Ornithology, DST-NRF Centre of Excellence, University of Cape Town, Rondebosch 7701, South Africa

* Correspondence:

Email: andremcferreira@cibio.up.pt

ABSTRACT

1. Individual identification is a crucial step to answer many questions in evolutionary biology and is mostly performed by marking animals with tags. Such methods are well established, but often make data collection and analyses time consuming, or limit the contexts in which data can be collected.
2. Recent computational advances, specifically deep learning, can help overcome the limitations of collecting large-scale data across contexts. However, one of the bottlenecks preventing the application of deep learning for individual identification is the need to collect and identify hundreds to thousands of individually-labelled pictures from which to train convolutional neural networks (CNNs).
3. Here, we describe procedures for automating the collection of training data, generating training datasets, and training CNNs to allow identification of individual birds. We apply our procedures to three small bird species, the sociable weaver *Philetairus socius,* the great tit *Parus major* and the zebra finch *Taeniopygia guttata*, representing both wild and captive contexts.
4. We first show how the collection of individually labelled images can be automated, allowing the construction of training datasets consisting of hundreds of images per individual. Second, we describe how to train a CNN to uniquely re-identify each individual in new images. Third, we illustrate the general applicability of CNNs for studies in animal biology by showing that trained CNNs can re-identify individual birds in images collected in contexts that differ from the ones originally used to train the CNNs. Finally, we present a potential solution to solve the issues of new incoming individuals.
5. Overall, our work demonstrates the feasibility of applying state-of-the-art deep learning tools for individual identification of birds, both in the lab and in the wild. These techniques are made possible by our approaches that allow efficient collection of training data. The ability to conduct individual recognition of birds without requiring external markers that can be visually identified by human observers represents a major advance over current methods.

Keywords: artificial intelligence, automated, convolutional neural networks, data collection, deep learning, individual identification

INTRODUCTION

In recent years, deep learning techniques, such as convolutional neural networks (CNNs), have caught the attention of ecologists. Such tools can automatize the analysis of various types of data, ranging from species abundance to behaviours, and from different sources such as pictures or audio recordings (reviewed in Christin, Hervet & Lecomte, 2019). CNNs are a class of deep neural networks that, contrary to other types of artificial intelligence methods that require hand-crafted feature extraction, automatically learn from the data the features that are optimal for solving a given classification problem (see Angermueller, Pärnamaa, Parts & Stegle, 2016; Christin et al., 2019; Jordan & Mitchell, 2015; LeCun, Bengio & Hinton, 2015 for a detailed introduction on deep learning). CNNs are thus particularly useful when many features for classification are needed.

In ecology, deep learning has been successfully and predominantly applied to identifying and counting animal or plant species from pictures. For example, Norouzzadeh et al. (2018) used a long term database of more than 3 million labelled pictures to train a CNN to automatically recognize 48 African animal species. This CNN can replace the need for manual identification in future studies, which is highly time consuming, thus promoting a more efficient data analysis pipeline. This, and other examples (e.g. Rzanny, Seeland, Wäldchen & Mäder, 2017; Tabak et al., 2019), highlight the potential for deep learning to help to increase sample sizes, and therefore help resolve many limitations in power for biological studies (e.g. Wang et al., 2018).

Beyond species recognition, one particularly promising application of CNNs is individual identification. Individual identification is crucial to many studies in ecology, behaviour and conservation (Clutton-Brock & Sheldon, 2010). The use of deep learning methods for individual identification has been the subject of extensive research in humans (e.g. Ranjan et al., 2018), where it has been extremely successful. More recently, a handful of studies have applied the similar methods to other animal species, allowing computers to individually-recognise primates (Deb et al., 2018; Schofield et al., 2019), pigs (Hansen et al., 2018), and elephants (Körschens, Barz & Denzler, 2018). However, the application of deep learning to smaller taxa, and specifically birds, remains unexplored.

In birds, manual examination of pictures or video recordings of visually marked populations is well established. For studies on both wild (i.e. free-ranging monitored populations) and captive animals, researchers often mark individuals with unique combinations of colour bands to facilitate observations in the field or, later, in recorded images. However, relying on humans for individual identification and data collection is extremely time-consuming (Weinstein, 2018). In the past decade, many studies have made use of automated animal-tracking devices (e.g. GPS) and sensor technologies (e.g. RFID) (reviewed in Krause et al., 2013). Such animal-borne tracking devices, however, often limit researchers to studying individuals in particular contexts. For many studies, obtaining visual records remains critically important. For example, studying parental care in birds requires video recordings to visually identify which birds are providing care to the chicks and how often they do it. Such data can, to some extent, be automated using PIT-tags and fitting RFID readers to a nest. However, this technology cannot record many additional, and important, pieces of information, such as the type of food that parents are bringing to the chicks or distinguishing the purpose of the visit (e.g. to feed the chicks or to engage in nest maintenance activities). Thus, a major advance over current methods would be to automatically identify individuals while keeping the versatility of the data and contexts that can be captured using pictures and video recordings.

Several methods for automatic individual identification and other data extraction from pictures and videos of animals have been developed previously. For instance, Pérez-Escudero, Vicente-Page, Hinz, Arganda & de Polavieja (2014) proposed a multi-tracking algorithm capable of following unmarked fish in captivity from video recordings (which was later improved using deep learning; Romero-Ferrero, Bergomi, Hinz, Heras, & de Polavieja, 2019). Other computer vision-based solutions rely on tags or marks to assist with computer tracking and individual identification (e.g. Alarcón-Nieto et al., 2018). To date, all these methods remain mostly limited to studying animals in captivity, either because they require standardized recording conditions (e.g. consistent background light, known number of individuals present in the recording) or the marks needed to assist individual identification are attached through gluing or using backpacks that are not suitable to be fitted to many animals, especially in the wild. Deep learning methods have the potential to overcome many of the limitations of the current automated methods, as they can identify individuals by relying only on the natural variation in appearance among individuals, while remaining tolerant to spurious variation arising from recording conditions.

A major challenge for the application of individual recognition using deep learning methods is the need of collecting extensive training data. Acquiring training data typically involves labelling images with the identity (or an attribute) of each individual. The amount of data required to train a CNN is expected to be proportionally dependent on the difficulty of the classification challenge, i.e. a bear and a bird would be easier to differentiate than two bears of the same species. Usually, CNNs that achieve large generalization capability need to be trained over thousands to millions of pictures (Marcus, 2018). Such large datasets are required because the aim of using a CNN is to generalize recognition from the specific data that the CNN has been exposed to during training. For example, if a CNN was trained to distinguish two bears of the same species with only pictures of the individuals lying down, it might be unable to identify those same individuals from new pictures taken when the animals are standing up. Additionally, if the pictures used for training were taken during a short period of time, it might lead the CNN to rely on superficial and temporary features for identification. For example, if pictures for training were taken when one of the individuals had a large wound or was going through moulting or shedding, it might result in a CNN that relies on those salient and temporary features, and thus perform badly when having to predict the identity of the individuals a few days later. Therefore, effectively making use of deep learning for individual identification, especially in the wild, requires new ways to collect training data that do not rely on individual manual image annotation.

When working in captivity settings, such large labelled image datasets can be easily collected by temporarily isolating the animals in enclosures separated from the rest of the group while filming or photographing them. However, such an approach is clearly not feasible for researchers working on wild populations, making collecting training data from wild animals much more challenging. For example, in birds, relying on human observers and colour rings, to photograph and manually label enough pictures to implement CNN for individual identification, would be extremely time-consuming. Furthermore, in longer-term studies, animals can change their appearance over time (e.g. changing from juvenile to adult plumage in birds) or new individuals may join the population (e.g. immigrants or recruited offspring). These cases require that the process of identifying individuals and labelling photos is routinely repeated. Therefore, relying on human observers for collecting labelled data in this type of systems might hinder the widespread implementation of deep learning techniques for individual identification, or restrict its application to short-term projects.

Here, we provide an efficient pipeline for collecting training data, both in captivity and in the wild, and we train CNNs for individual re-identification (i.e. machine recognition of a previously known set of individuals). We demonstrate the feasibility of our approaches using data from two wild populations of birds of two different species, the sociable weaver *Philetairus socius* and the great tit *Parus major,* and a population of captive zebra finches *Taeniopygia guttata*. We then show that CNNs trained on these species can successfully re-identify individuals across a range of different contexts.

We start by 1) focusing on the problem of efficiently collecting large training datasets. We provide simple and automated methods for collecting a very large number of labelled pictures by using low-cost cameras that can be programmed to take labelled pictures of birds. In captivity, we achieve this by temporarily isolating target individuals, and taking pictures using low-cost cameras. In the wild, we describe a solution using low-cost RFIDs and low-cost cameras that are programmed to take labelled picture when PIT-tagged birds land on an RFID-equipped feeder. We then 2) provide details of the steps involved with data pre-processing and the training of an adequate CNN. We further describe approaches for augmenting our training datasets using algorithms that add noise and make modifications to the original images. Next, we 3) evaluate the generalization performance of our CNNs to data collected in other contexts by evaluating the ability of our models to predict the identity of the birds in pictures collected using different cameras and in contexts that differ from the ones used for collecting the training datasets. Finally, we 4) present a very simple approach to address the problems arising from the arrival of new and unmarked individuals to the population.

METHODS:

Study populations:

We collected pictures from a population of sociable weavers at Benfontein Nature Reserve in Kimberley, South Africa, and a population of great tits, from a population in Möggingen, southern Germany. For both species, individuals were fitted with PIT-tags as nestlings, or when trapped in mist-nets as adults, and were habituated to artificial feeders that are fitted with RFID antennas, as part of on-going studies in these populations. We also collected data from a captive population of zebra finches housed in Möggingen, southern Germany. Birds from this population were being kept in indoor cages in pairs and small flocks.

Collecting training data:

In all three species, we collected pictures using Raspberry Pi cameras. The methods to automatically label the pictures differed between the wild (sociable weavers and great tits) and captive (zebra finches) populations. We start by explain the two different data collection pipelines.

Training data collection in the wild:

The collection of labelled pictures in the wild was automated by combining RFID technology (Priority1Design, Australia), single-board computers (Raspberry Pi), Pi cameras, and artificial feeders. We fitted RFID antenna to small perches placed in front of bird feeders filled with seeds (Fig. 1a, b and c). The RFID data logger was then directly connected to a Raspberry Pi (detailed explanation of the developed setup is available at github.com/AndreCFerreira/Bird_individualID) which had a Pi camera (we used Pi camera V1 5mp and V2 8mp). When the RFID data logger detected a bird, it sent the individual's PIT-tag code to the Raspberry Pi, which was programmed to then take a picture. Because birds often spend some time on the feeder, we programmed the Raspberry PI to take a picture every 2 seconds while the bird remained present. This interval was introduced in order to efficiently collect data while avoid having near-identical frames of the same bird as having too many near-identical pictures could increase the overfitting of the CNN, i.e. the risk of the model "memorizing" the pictures instead of learning features that are key for recognizing the individuals and thus jeopardize the generalization capability of the models (see "Convolutional neural networks" section). Each picture file was automatically labelled with the bird identity, known from the RFID logger and the time of shooting in the filename. Training data collection was therefore automatized by linking the identity of the bird perching on the antenna while feeding to its pictures, without any need for human manual identification and annotation. When multiple birds perched on the feeder at the same time, it was not possible to determine which of the birds activated the RFID system. Pictures that contain more than one bird were thus automatically excluded (see "Data pre-processing" section).

For the sociable weaver population, we placed three PI cameras and three feeders on the ground about two meters apart from each other. For the great tit population we used one PI camera fitted to one feeder hanging on a tree branch. The cameras were positioned to take a picture from top perspective to enable to photograph both the back and wing feathers (Fig. 1b, c). The birds' back was chosen as the distinctive mark since it is the body part that is most easily observed and recorded in multiple contexts (e.g. when perching at the feeders or building at the nest), making it a very versatile mark for applying an image classification algorithm in other contexts. For the sociable weaver population, we collected images for 15 days during November and December 2018. For the great tit population, we collected images over seven days during the last two weeks of August 2019.

Training data collection in captivity:

We temporarily divided cages into equally-sized partitions with a net, allowing us to take pictures from individual birds without completely socially isolating them. We collected data from 10 zebra finches (five males and five females). We placed two Raspberry Pi cameras on the roof of each partition to photograph (every two seconds) the birds sitting on the wooden perches (Fig. 1d). Each bird was recorded for four hours. Since we knew which Raspberry Pi photographed which bird, we avoided the need to manually link the identity of the birds to the pictures.

Data pre-processing:

To efficiently train a CNN, the regions in the pictures corresponding to the birds should be extracted from the background (third step of Fig. 2). A Mask R-CNN (He, Gkioxari, Dollár & Girshick, 2017) was used to automatically localize and crop the bird in the pictures. For the sociable weavers, we used a Mask R-CNN model that had been trained on Microsoft COCO (Lin et al., 2014). Microsoft COCO is a generalist dataset which includes pictures of birds and therefore is able to localize the sociable weavers in the pictures (see github.com/AndreCFerreira/Bird_individualID for details). Because the sociable weaver population was colour-banded, and these were partially visible in some of the cropped pictures, we manually removed any visible colour bands from the testing data (see “Testing models” section) to ensure that colour bands were not used for individual identification by the model.

As the Mask R-CNN model performed poorly for the great tits and zebra finches, we re-trained the model by adding a new category (zebra finch or great tit, making a different model for each species) using pictures in which the region corresponding to the bird was manually delimited using “VGG Image Annotator” software (Dutta & Zisserman, 2019). Since manually labelling the regions of interest is time consuming, we started by training the model for 10 epochs (i.e. passing the entire dataset through the neural network 10 times) with 200 manually labelled pictures. If the model was found to perform badly, additional pictures were manually labelled and added it to the training dataset. This process was repeated until a satisfactory performance was achieved. For the great tits, we needed 500 pictures in the training data and 125 for validation (see “Convolutional neural networks” section below for explanation on training and validation datasets). The zebra finch data required 400 pictures for training and 100 for validation.

For the sociable weavers and the great tits, if the Mask R-CNN identified more than one bird perching simultaneous at the RFID antenna, we automatically excluded that image. We detected a total of 35 sociable weavers at the RFIDs antennas. Of these, 30 individuals with more than 350 pictures were used to train the classifier. In the great tit population, 77 birds were photographed, of which 10 had more than 350 pictures. These 10 individuals were used to train a CNN for each of the species. The remaining five sociable weavers and 67 great tits (with less than 350 pictures) were used to address the issue of working in open areas where new individuals can constantly be recruited to the study population (see section “New birds” below). For the zebra finches we used all 10 individuals as our setup resulted in more than 2000 pictures for each bird.

Convolutional neural networks:

Training a CNN requires both a training and a validation dataset. The training dataset is the set of samples that the neural network repeatedly uses to learn how to classify the input images into different classes (in our case, different individuals). The validation dataset is an independent set of samples that is used to compute the accuracy and loss (estimation of the error during training) of the model. This validation dataset is used to assess the learning progress of the neural network. As the network never trains on or sees the validation data, this validation dataset can indicate if the model is overfitting the training data and not learning features that are key for recognizing the individuals. It is generally difficult to anticipate the minimum number of images needed from each individual to obtain high performance for individual recognition. As a compromise between the number of birds that we could include in our study and the number of images per bird (i.e. to avoid generating an excessively imbalanced dataset), we aimed to use 1000 images per bird: 900 images for the training dataset and 100 images for the validation dataset. Training a deep learning model with an imbalanced training dataset (i.e. when the different classes, here the individuals, have different number of training pictures) can result in the over-generalization for the classes in majority due to its increased prior probability. For instance, a naïve classifier for a binary classification task for a dataset in which the ratio of the minority class to the majority class is 1:100 will have 99% accuracy if it simply learns to always output one result: the majority class. As a consequence of this, data containing minority classes (in our case birds with fewer images) are more likely to be misclassified than those belonging to the majority classes (Johnson & Khoshgoftaar, 2019). One countermeasure against class imbalance is oversampling, which consists of creating copies of the training data from the less sampled classes.

We applied limited oversampling to our training dataset only. For 9 sociable weavers and 6 great tits for which we did not have 1000 images, we first selected 100 images for the validation dataset and then duplicated (through oversampling) the remaining pictures until

900 images were available for the training dataset (Buda, Maki & Mazurowski, 2018). Oversampling was therefore restricted to the training dataset and not applied to the validation dataset in order to avoid overestimating the model's learning progress. For both species, in order to limit overfitting caused by having very similar pictures in the training and validation datasets, we used images from different days in our training and validation datasets. In total, we constructed a dataset of sociable weavers containing 27038 unique images of 30 individuals, or 901±173 (mean±SD) per bird and a dataset of great tits containing 7605 unique images of 10 individuals, 761±223 (mean±SD) per bird.

Working on the captive zebra finches, we could easily collect many images per bird. However, the problem of collecting data of animals that are in confined enclosures is that a significant number of pictures could potentially be near-identical, such as if an individual stays motionless for long periods of time. In our case, all birds were generally active and visited all the places in their cage (i.e. all wooden perches, floor, water and food plates). Nevertheless, to avoid potential overestimation of the model's accuracy, we used the images collected when the birds were in different partitions for training and validation datasets. Additionally, to create a diverse set of validation pictures, we used a structural-similarity index measure (SSIM; Wang, Bovik, Sheikh & Simoncelli, 2004) to create a dataset with maximised pairwise dissimilarity among images (following a similar procedure as Hansen et al., 2018 for a pig dataset). We started by randomly selecting an image to include in the validation dataset. We then randomly sampled images and computed the SSIM between the new image and those already in the validation dataset. If the SSIM value was smaller than a threshold, these new pictures were included in the validation dataset. This process was repeated by sequentially comparing a new picture to all the ones already in the validation dataset until we reached 160 images per bird. The threshold value used (0.55) was empirically determined by trying different values and looking at the resulting datasets. For the training dataset, 1600 images of each zebra finch were randomly selected without filtering for near-identical images. All birds had at least 1600 images, except for one that had

1197 for which oversampling was used by creating duplicates of 403 randomly sampled images.

We used the VGG19 convolutional neural network architecture (Simonyan & Zisserman, 2014) and initialized the model with the weights of a network pre-trained on the ImageNet dataset (a dataset with more than 14 million pictures and 20000 classes, Deng et al., 2009). The main idea behind using networks pre-trained on other datasets is that features (such as colour or texture) that are important to distinguish multiple objects could also be useful to distinguish between individual birds. When using transfer learning, the bottom layers of the network can be frozen in order to mitigate overfitting, this is especially important when the training datasets are small. However, as freezing the layers prevent them from update their weights during the training process (and therefore could prevent the model from learning key features for performing the classification task) and considering the size of our training datasets, we decided to train the models without freezing any of the layers of the network. The fully connected part of the VGG19 CNN network (i.e. the classifier part) was replaced by layers with random weights that fit our particular task of interest and the corresponding number of classes (i.e. number of different individuals; Supplementary Fig. S1).

To further increase our training sample, we then used a data augmentation procedure. This procedure consists of artificially increasing the sample size by applying transformations to an existing set of samples. Using the data generator available in Keras (Chollet, 2015), we randomly rotated (from 0 to 40º) and zoomed (zoom range of 0.2) images of all species. We additionally applied horizontal and vertical flips to the great tits and zebra finches populations, as contrary to the sociable weavers, these birds could be photographed from any orientation (as they perched all around the RFID antenna or the cage perch their bodies can be facing different directions). These transformations were applied randomly to every single picture in the dataset as the Keras generator does not provide the original images directly to the model during training. Instead, only augmented images are provided to the model in each epoch, but since transformations are performed randomly, both modified images and close reproductions of the original images (i.e. those with almost no augmentation) are provided during training.

One dropout layer was added just before the first dense layer (see github.com/AndreCFerreira/Bird_individualID and Supplementary Fig. S1 for details on the network architecture). Dropout layers are used to limit overfitting by randomly ignoring units of the CNN (i.e. neurons) during the training process (see Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov, 2014 for details on dropout). For the sociable weavers and the zebra finches, the dropout layer had a value of 0.5, while for the great tits it was reduced to 0.2 (i.e. less units are being ignored in order to facilitate the training process) as the model did not improve the accuracy from a random guess for 10 epochs when the dropout was at an initial value of 0.5. We used a softmax activation function for the classifier and ADAM optimizer (Kingma & Ba, 2014) with a learning rate of 1^{e-5}. A batch size of eight (i.e. eight pictures are being provided to the model each time) was used since it has been shown that small batch sizes improve models' generalization capability (Masters & Luschi, 2018). If there was no decrease in loss (i.e. measure of the difference between the predicted output and the actual output) for more than 10 consecutive epochs, we stopped training, and then retrained the model that achieved the lowest loss with a SGD optimizer and a learning rate 10 times smaller until there was no further decrease in the loss for more than 10 consecutive epochs. All pictures were normalized by dividing the arrays by 255 (0 to 1 normalization). All analyses were conducted with python 3.7 using Keras tensorflow 1.9 on an Nvidia RTX 2070 GPU.

In the case of the sociable weavers (which was the species that we used when initially exploring our approaches), even though our model achieved ca. 90% accuracy with the validation dataset, the accuracy was significantly lower when generalizing to other contexts (see "Testing models" and "Results" sections). We suspected that such differences could be due to the lower quality of images collected in those other contexts (with different cameras, capture distances and conditions; see "Testing models" section). To account for this possibility, we trained a model using the same setting parameters that yielded the best results, and applying further transformation. In order to simulate the lower quality of the pictures taken in other contexts, we applied Gaussian blur, motion blur, Gaussian noise, resizing transformations and a random combination of two of these four transformations (see github.com/AndreCFerreira/Bird_individualID for details on the transformations applied to the images) to each of the images in the dataset used to train the models (Fig. 3). The idea is that even if the overall quality of the pictures in the dataset used for training slightly differs from pictures which are of interest for a research question, this training dataset can be transformed in order to be more similar to the pictures collected in distinct contexts for which the classifier could be applied on. Blur and noise transformations were not used for the great tits and zebra finches as there were no differences in the overall quality of the pictures used for training and for testing the model generalization capability (see “Testing models” section).

Testing models

To test the efficiency of our models, we collected images of birds in different viewing perspectives, using different cameras, and across different contexts than the original feeding station setup. The aim was to evaluate the ability of our trained CNN to identify individuals in different experiments and contexts, and to verify that the models were not overfitting the training data.

For the sociable weavers, we used four different setups for testing. We filmed birds feeding in the same plastic RFID feeders but recorded using a Sony handycam (rather than Raspberry Pi camera), from two different perspectives: 1) close (ca. 30 cm from the feeder, 95 images of 26 birds 3.65 ± 0.68 (mean ± SD; Fig. 4a) and 2) and far (ca. 100 cm from the feeder, 71 images of 21 birds 3.43 ± 0.58; Fig. 4b). In addition, a plastic round feeder with seeds was positioned on the floor to record both from 3) a ground perspective (90 images of 28 birds 3.21 ± 1.21; Fig. 4c) and 4) a top perspective (83 images of 25 birds 3.32 ± 1.01; Fig. 4d). The birds were manually cropped out from pictures using imageJ (Schneider, Rasband & Eliceiri, 2012) and individually identified using their colour rings. The colour rings were then erased directly from the image to guarantee that the model did not use them for identification. Videos were recorded within the same time window as the training pictures collection and we aimed to extract five non-identical frames per bird in which the back was fully visible. Unfortunately, this was not always possible for all birds as not all of them were present or recorded long enough in these testing videos and therefore the sample size for each perspective differs.

For the great tits, we recorded the birds feeding in a table from a top perspective with a Raspberry Pi camera (Fig. 4e). Since these birds had no colour ring or any mark for visual identification, we identified them using their PIT-tags by placing seeds on top of a RFID antenna that was on a feeding platform. Birds were recorded feeding on the table for 3 days, but 4 out of the 10 birds used in the training dataset did not use this new feeding spot. In all, 94 pictures were taken but the number of pictures collected at this setup varied greatly between birds (from 2 to 38 pictures, mean: 15.70±11.30SD). As a result, we did not attempt to make a balanced dataset and, therefore, used all the 94 pictures collected at this new feeding setup.

For the zebra finches, we did not have a second setup that differed from the one used to collect the pictures to train a CNN and that could be used for testing the CNN generalization. Instead, we ran an additional trial which consisted of recording the birds together to see how well the model would predict the identity of each individual when they are in small groups interacting with each other (Fig. 4f). Since these birds did not have any visual tags and it was not possible to distinguish them when in group, we used one flock of three birds and another flock of two birds for each sex. This allows us to estimate the model's accuracy by calculating the number of times that the CNN wrongly attributed the identity of a bird as being an individual that was not actually present in that flock. In order to avoid near-identical pictures, the same procedure as for the validation dataset to select 160 pictures from each trial was used.

New birds:

In the wild, it is common for new individuals to join a population during the course of a study. These new individuals may challenge the performance of a CNN, because the model outputs a vector from a softmax layer that indicates probabilities of presence for every individual present during training, with the sum of these probabilities being one (see “classification” stage in Fig. 2). In order to study this potential issue, we used the already trained CNNs from the subset of identities we had to predict the identity of birds that were not included in the training datasets. For the sociable weavers, we had a scenario in which a CNN was trained to identify a relatively large number of individuals (30) was then exposed to a small number of new individuals (5). For the great tits, we had the opposite scenario in which a CNN that was trained for a small number of individuals (10) was then exposed to a large number of new individuals (67). For the sociable weavers, we selected 50 pictures of each of the five birds (a total of 250) that were not in the training dataset and 250 random images from the pool of birds that were included in the training data. For the great tits, we selected 250 random images from the pool of 67 individuals that were not in the training dataset, and kept a random set of 250 images from the birds in the training data. We limited the number of pictures from the same individual to a maximum of eight (3.91± 1.67 mean±SD) in order to keep a large number of different individuals in this dataset (64 out of the 67 individuals were used). Shannon’s entropy (Shannon, 1948) of each of the distributions was calculated from the classification (softmax) output to empirically determine a confidence threshold to consider a bird as part of the training dataset.

RESULTS

CNN

Sociable weavers

The model was able to achieve an accuracy of 92.4% (Table 1) after training for 21 epochs (ca. 360min of training). When the model was used to predict the identity in four other contexts, it appears that the accuracy of top perspective's context was lower (67.5% for the plate top Table 1). After adding blur and noise to the training images, the model achieved a validation accuracy of 90.3%, while successfully increasing the accuracy from the top perspective to 91.6% (Table 1).

Great tits

The model reached 90.0% accuracy after training for 32 epochs (ca. 105min). When using the pictures from the top perspective recording the birds on the table the model correctly predicted the identity of the birds in 85.1% of the pictures.

Zebra finches

The model reached 87.0% accuracy after training for 11 epochs (ca. 150min), and obtained similar accuracies for males and females (85% for males, 88.9% for females). When using the trained model to predict the identity of the birds when they were in small groups the model correctly predicted the identity of a bird present in that group in 93.6% of the time.

New birds

The entropy of the softmax outputs (i.e. probabilities) was smaller when predicting the identity of birds present in the training dataset, compared to when predicting the identity of new birds (Fig. 5). This is due to the fact that when predicting the identity of a bird from the training dataset, there is usually one that stands out with very high probability (thus successfully indicating the bird's identity) and the remaining probabilities are very low (other birds' identities). In contrast, when predicting the identity of a new bird, the probabilities were usually more equally distributed across all classes, all with low values.

For the sociable weavers, 90% of entropies were below 0.75 when predicting the identity of birds from the training dataset and only 17% of them were under this value when predicting the identity of new birds. This means that with this 0.75 threshold there is a 17% chance that a new bird will be erroneously classified as one of the birds of the training dataset. A value of 17% should be acceptable if new individuals are not common (both in number of different new individuals and in the frequency of appearance). In order to reduce the probability of identifying a new sociable weaver as a bird present in the training dataset to less than 5%, a confidence threshold for the entropies would have to be set to 0.018. However this would result in discarding 36% of the images of the sociable weavers present in the training dataset.

For the great tits scenario, in which the appearance of new birds is frequent, defining a simple threshold that differentiate new birds from the birds already present in the training dataset would not be enough as there is a too much overlap between the birds in the training and the new birds' entropy. For example, 90% of the entropies are below 0.8 when predicting the identity of birds that are present in the training dataset. However 62% of the entropies for the birds not present in the training dataset are also below this value. Under this scenario, reducing the probability of identify an new individual as a bird present in the training dataset to less than 5% would require to set a confidence threshold for the entropies of 0.002 which would result in discarding 77% of the images of birds present in the training dataset.

DISCUSSION

Deep learning has the potential to revolutionize the way in which researchers identify individuals. Here, we propose a practical way of collecting large labelled datasets, which is currently the main bottleneck preventing the application of deep learning for individual identification in animals (Schneider, Taylor, Linquist & Kremer, 2018). We also show the steps required to train a classifier for individual re-identification. To our knowledge, this is the first successful attempt of performing such an individual recognition in small birds. Using data collected with automatized procedures, CNNs proved to be effective for re-identifying known individuals in three different bird species, including two species that are among the most commonly used models in the field of behavioural ecology (great tits and zebra finches). Our results therefore clearly highlight the potential of applying CNN to a vast range of research projects. Furthermore, we found that our trained CNNs were generalisable, meaning that the rate of successful re-identification remained high across different recording contexts. This is particularly relevant as researchers often interested in collecting data in contexts that are challenging, from parental behaviour at the nest to dominance interactions away from artificial feeders. However, we also show that the models' performance can be reduced when new individuals join the population, especially when new individuals are common.

The first critical step when deciding whether to implement a deep learning approach for a given study is to guarantee that enough training data can be collected to train a model. Our data from two wild populations showed that we can rely on RFID technology to gather large amounts of automatically labelled data. Since this technology is now widely used for research on birds (e.g. Aplin et al., 2015), we believe that the proposed method for automatizing data collection for deep learning applications could be easily and rapidly implemented in a large number of research programs. The advantage that deep learning would offer is to be able to collect data from much more general contexts, away from a feeding context (which is usually where RFID readers are placed). Furthermore, the method could be easily extended to other animals and other identification techniques. The main idea is to develop a framework in which the same individuals can be repeatedly encountered, at which time the images that are recorded are automatically labelled. For example, GPS (e.g., Weerd et al., 2015) or proximity tags technology (e.g., Levin, Zonana, Burt & Safran, 2015) could also be used in combination with camera traps to collect training data. Even with non-electronic tags, it should be possible to design setups to photograph animals automatically, such as by isolating the animals as we showed here with the zebra finches. With the popularization of imaging and sensor technologies, we believe that efficiently collecting a large amount of data should no longer represent a bottleneck preventing the application of deep learning methods such as CNN.

The most powerful aspect of CNNs is that they can provide a generalised identification solution. However, the capacity for a CNN to work effectively across contexts will be affected by variation in the recording conditions, for example due to light intensity, shadow or characteristics inherent to the recording quality. One solution to this is to ensure that the training dataset contains sufficient variation to capture the broad range of contexts that the CNN is required for. Photographing the animals across different times of the day and in different days provides the CNN with a very diverse training dataset making the CNN invariant to such variations. Furthermore, we show here that if the conditions for training are slightly different from the recording conditions in which the CNN is going to be applied, it is possible to artificially modify the pictures used for training in order to simulate the conditions under which the pictures of the context of interest will be taken. Specifically, we used blur and noise transformations in the sociable weaver dataset to improve the generalization capability of our model, as the testing images had a lower quality than the training images. This confirms that using artificially degraded training pictures can be used to improve CNN generalization capability (e.g. Vasiljevic, Chakrabarti & Shakhnarovich, 2016). Other transformations could potentially be applied on the training dataset. Such transformations should consider the type of images on which the model will be used. For example, if illumination conditions of the training pictures are different from the context of interest, brightness and contrasts transformations could be applied to the training data in order to make the CNN light invariant. This generalization capability is an important novelty of this study compared to previous work on small-animal tracking using computer vision, which have been restricted to standardized conditions (e.g. Pérez-Escudero et al., 2014) that are not easily satisfied when working with wild animal populations.

Besides the recording conditions, it is also important to consider how tags used for human identification could artificially increase the accuracy of the models. For example, here the sociable weavers had 3 coloured bands and a metal ring in their legs (two in each leg) that form a unique colour combo code. The Mask R-CNN trained on Microsoft COCO dataset used here to extract the birds from the pictures resulted in a dataset with 36% of the pictures containing at least one of the 3 colour bands partially visible, whereas the full colour code was almost never visible (fewer than 1% of the pictures). Since the majority of the pictures did not have any colour band visible, and 3 colour rings are needed to correctly identify the individuals (there are large overlaps between the colour bands, e.g. 6 birds had an identically-positioned black band), we are confident that no additional effort would have been needed to remove the colour bands from the training or validation datasets. We confirmed this by manually removing the colour bands from all testing pictures, and finding that the model maintained the same accuracy as the validation dataset (ca. 90%). However, in situations in which colour bands might represent a real issue, a Mask R-CNN could be specifically trained to extract the bodies of the birds without their legs.

Another major challenge to the applicability of CNNs is dealing with temporal changes in the appearance of individuals. For research questions that do not need long time windows of data collection or that are conducted on species that maintain their appearance with great consistency, collecting training data within a short-period of time might be sufficient to develop a robust algorithm for individual identification. However, for longer-term studies, or when working with species that have the potential to change their appearance (e.g. moulting in birds), temporal changes in appearance constitutes a potentially serious limitation. The problem of long-term application of neural network algorithms has been studied in the context of place recognition (e.g. streets recognitions; Gomez-Ojeda, Lopez-Antequera, Petkov, & Gonzalez-Jimenez, 2015); however, to our knowledge, there is still no study addressing the impact of changes in appearance in animals in deep learning-based solutions. Currently, we do not know how CNNs would perform over long periods of time.

Solutions that could be explored include training data collected during long periods of time or targeting specific parts (e.g. excluding the wing feathers and considering only the top part of the back, or other body parts of the birds such as the flank or the bib) of the birds. These could make the CNN appearance-invariant by learning more conservative features of the birds that are kept across time (even through moulting events). In order to fully address the problem and the potential solutions, images of birds collected over longer periods of time and from multiple body parts are needed. At present, such datasets are not available. However, the automatization of training data collection is an immediate and effective solution, i.e. it now feasible to continuously collect training pictures and routinely re-train a CNN using updated training data.

The arrival of new individuals to the study population is another challenge that needs to be carefully addressed. If these new birds are marked with a PIT-tag, the CNN could be updated similarly to the problem of changes in appearance discussed above. However, in many cases new individuals will not be marked. Such a problem fits in the anomaly (Chandola, Banerjee & Kumar, 2009) and novelty (Pimentel, Clifton, Clifton & Tarassenko, 2014) detection domain. Here, we explored a simple approach involving investigating the entropy of classification probabilities. Our solution appears useful if the CNN was trained on a relatively large number of individuals and if immigrants are uncommon in the population, like in the sociable weaver example. However, for some studies, such conditions might not be met and, as it was the case of the great tit scenario, where we had a low number of individuals in the training dataset and observed a large number of new birds. Nevertheless, the identification accuracy of a CNN should also be considered from a post-detection analysis perspective. While some studies will benefit from maximise the number of identifications made, in other studies it may be more costly to have misidentified individuals For example, misidentifications are very costly when construction social networks (Davis, Crofoot & Farine, 2018), while at the same time social networks are very identification hungry (Farine & Strandburg-Peshkin, 2015). Thus, exploration of the entropy distribution and other approaches, and subsequent trade-offs, should be considered. In addition, the error rate might be also reduced through post-processing. For example, if the identification is based on a collection of frames (e.g. images extracted from a short video recording of the animal) instead of single image, then the sequence of detections (and assignment probabilities) can be quantified over subsequent frames, and the detection can be kept or discarded depending on the overall confidence in the sequence of detected identification.

The field of deep learning progresses rapidly and almost continuously provides solutions to seemingly challenging problems. However, this is facilitated by the existence of large and freely availed databases, which are used to try different approaches for a wide range of classification problems. For example, the ImageNet database (Deng et al., 2009) has been used numerous times to create algorithms for object recognition. The Labelled Faces in the Wilde (LFW) dataset (Huang, Mattar, Berg & Learned-Miller, 2008) contains thousands of pictures of human faces to development algorithms for human face recognition and identification. The nordland dataset (Sünderhauf, Neubert & Protzel, 2013) contains footage of more than 700km of northern Norway railroad recorded in different seasons (summer, winter, spring and fall) and has been used to address the problem of place recognition under severe environmental changes. Biologists aiming at taking advantage of the potential of deep learning will also benefit from assembling large datasets of labelled pictures containing many individuals, taken across different contexts and across different life stages. By making our dataset freely-available, we provide the foundations for continued development of more reliable algorithms that are able to cope with the challenges presented here, among others.

Having large datasets will allow optimizing performance of CNNs as well as identifying the relative performance of alternative solutions. Other network architectures (e.g. ResNet; He, Zhang, Ren & Sun, 2016) and different hyper-parameters settings (e.g. learning rate) than the ones used here can yield different, and potentially improved, results. Other deep learning methods approaches could also be explored and applied not only to closed-set identification problems (as we did here) but also to verification and open-set identification. For example

Siamese neural networks (Varior, Haloi & Wang, 2016) and triplet loss based methods (Schroff, Kalenichenko & Philbin 2015) are able to make pairwise comparison of two different images and output if the different images belong to the same individual or not, which could help solve the issue of the introduction of new individuals to the population and obtain higher overall performance. There are also other pre-processing steps that can greatly improve the model training and reduce the number of images needed. For example, image alignment (e.g. Deb et al., 2018; Lopes, de Aguiar, De Souza, & Oliveira-Santos, 2017) can be used to decrease variation in the birds' pose. Training an algorithm for individual recognition encompasses a great deal of trial and error, and different systems will present different challenges, but also opens up many new opportunities. Comparison of the performance of different methods for individual recognition in birds should therefore be the scope of intense research once sufficient individually labelled dataset becomes available.

We hope that our work will motivate other researchers to start exploring the possibility of using deep learning for individual identification in their model species. More work is needed to address the constraints of working with birds both in the wild and in captivity (namely moulting and introduction of new individuals). However, the ability to move beyond visual marks and manual video coding will revolutionise our approach to addressing biological questions. Importantly, it will allow researchers to expand their sample sizes, thereby providing more power to test hypotheses. Finally, it will open up opportunities to address questions that previously were not tractable.

ACKNOWLEDGEMENTS

Data collection on the sociable weaver population would have not been possible without the contribution of several people working in the field, in particular those who contributed to operate the RFID stations and conduct the annual sociable weavers captures: António Vieira, Rita Fortuna, Pietro D'Amelio, Cecile Vansteenberghe, Franck Theron, Annick Lucas, Sam Perret, and several other volunteers. De Beers Consolidated Mines gave us permission to work at Benfontein Reserve. We also thank Gustavo Alarcón-Nieto and Adriana Maldonado-Chaparro for the assistance with the material needed to collect pictures of the great tits and the zebra finches. Data collection for the sociable weaver data was supported by funding from the FitzPatrick Institute of African Ornithology (DST-NRF Centre of Excellence) at the University of Cape Town (South Africa), FCT (Portugal) through grants IF/01411/2014/CP1256/CT0007 and PTDC/BIA-EVF/5249/2014 to RC and the French ANR (Projects ANR- 15-CE32-0012-02 and ANR 19-CE02-0014-02) to CD. This work was conducted under the CNRS-CIBIO Laboratoire International Associé (LIA). ACF was funded by FCT SFRH/BD/122106/2016. FR was funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Individual Call - CEECIND/01970/2017. This work benefited from grants awarded to DRF by the Deutsche Forschungsgemeinschaft (DFG grant FA 1402/4-1) and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 850859). DRF and HBB received additional funding by the Max Planck Society and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 – 422037984.

AUTHORS' CONTRIBUTIONS

ACF, LRS, CD and JPR had the idea of applying deep learning for individual identification in the sociable weaver population and DRF had the idea of applying it to the zebra finch and great tit populations. ACF and LRS developed the RFID and Raspberry Pi based method for automated training data collection. LRS analysed the sociable weaver videos for testing the model generalization capability. RC and CD provided all the required funding, material and access to the individually marked sociable weaver population and DRF to the great tit and zebra finch populations. ACF, HBB and DRF developed the setup to collect pictures of the zebra finches. ACF, HBB collected the data of the zebra finches. ACF collected the data for the sociable weaver and the great tit populations. ACF led the statistical analysis and data pre-processing assisted by FR and JPR. ACF wrote the first draft of the manuscript. All authors contributed to editing and revising the final manuscript.

DATA ACCESSIBILITY

All scripts and data for reproducing the entire contents of this article are available at https://github.com/AndreCFerreira/Bird_individualID.

REFERENCES:

Alarcón-Nieto, G., Graving, J. M., Klarevas-Irby, J. A., Maldonado-Chaparro, A. A., Mueller, I., & Farine, D. R. (2018) An automated barcode tracking system for behavioural studies in birds. *Methods in Ecology and Evolution*, 9(6), 1536-1547.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016) Deep learning for computational biology. *Molecular systems biology*, 12(7), 878.

Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015) Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, 518, 538-541.

Buda, M., Maki, A., & Mazurowski, M. A. (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249-259.

Chandola, V., Banerjee, A., & Kumar, V. (2009) Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15.

Christin S, Hervet É, Lecomte N. (2019) Applications for deep learning in ecology. *Methods in Ecology Evolution*. 00:1–13

Clutton-Brock, T., & Sheldon, B. C. (2010) Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends in ecology & evolution*, *25*(10), 562-573.

Chollet, F. (2015) Keras: The Python Deep Learning Library. *Astrophysics Source Code Library*. Available online: https://keras.io/

Davis, G. H., Crofoot, M. C., & Farine, D. R. (2018) Estimating the robustness and uncertainty of animal social networks using different observational methods. *Animal Behaviour*, 141, 29-44.

Deb, D., Wiper, S., Russo, A., Gong, S., Shi, Y., Tymoszek, C., & Jain, A. (2018) Face recognition: Primates in the wild. arXiv preprint arXiv:1804.08790

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 248–255). IEEE.

Dutta, A., & Zisserman, A. (2019) The VGG image annotator (VIA). arXiv preprint arXiv:1904.10699.

Farine, D. R., & Strandburg-Peshkin, A. (2015) Estimating uncertainty and reliability of social network data using Bayesian inference. *Royal Society open science*, 2(9), 150367.

Gomez-Ojeda, R., Lopez-Antequera, M., Petkov, N., & Gonzalez-Jimenez, J. (2015) Training a convolutional neural network for appearance-invariant place recognition. arXiv preprint arXiv:1505.07428.

Hansen, M. F., Smith, M. L., Smith, L. N., Salter, M. G., Baxter, E. M., Farish, M., & Grieve, B. (2018) Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98, 145-152.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017) Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments.

Johnson, J. M., & Khoshgoftaar, T. M. (2019) Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 27.

Jordan, M. I., & Mitchell, T. M. (2015) Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

LeCun, Y., Bengio, Y., & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436.

Levin, I. I., Zonana, D. M., Burt, J. M., & Safran, R. J. (2015) Performance of encounternet tags: field tests of miniaturized proximity loggers for use on small birds. *PloS one*, 10(9), e0137242.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014) Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), European conference on computer vision (pp. 740–755). Cham: Springer

Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61, 610-628.

Kingma, D. P., & Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Körschens, M., Barz, B., & Denzler, J. (2018) Towards automatic identification of elephants in the wild. arXiv preprint arXiv:1812.04418.

Krause, J., Krause, S., Arlinghaus, R., Psorakis, I., Roberts, S., & Rutz, C. (2013). Reality mining of animal social systems. *Trends in ecology & evolution*, 28(9), 541-551.

Marcus, G. (2018) Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

Masters, D., & Luschi, C. (2018) Revisiting small batch training for deep neural networks. arXiv preprint arXiv:1804.07612.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716-E5725.

Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S., & De Polavieja, G. G. (2014) idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nature methods*, 11(7), 743.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014) A review of novelty detection. *Signal Processing*, 99, 215-249.

Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J. C., Patel, V. M., ... Chellappa, R. (2018) Deep learning for understanding faces: Machines may be just as good, or better, than humans. IEEE Signal Processing Magazine, 35(1), 66-83.

Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J., & de Polavieja, G. G. (2019) idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nature methods*, 16(2), 179.

Rzanny, M., Seeland, M., Wäldchen, J., & Mäder, P. (2017) Acquiring and preprocessing leaf images for automated plant identification: understanding the tradeoff between effort and information gain. *Plant methods*, *13*(1), 97.

Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012) NIH Image to ImageJ: 25 years of image analysis, *Nature methods* 9(7): 671-675

Schneider, S., Taylor, G. W., Linquist, S., & Kremer, S. C. (2019) Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4), 461-470.

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019) Chimpanzee face recognition from videos in the wild using deep learning. *Science advances*, 5(9), eaaw0736.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015) Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition

Shannon, C. E. (1948) A mathematical theory of communication. *Bell system technical journal*, 27(3), 379-423. (pp. 815-823).

Simonyan, K., & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.

Sünderhauf, N., Neubert, P., & Protzel, P. (2013) Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA).

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., ... Teton, B. (2019) Machine learning to classify animal species in camera trap images: applications in ecology. *Methods in Ecology and Evolution*, *10*(4), 585-590.

Varior, R. R., Haloi, M., & Wang, G. (2016) Gated siamese convolutional neural network architecture for human re-identification. In European conference on computer vision (pp. 791-808). Springer, Cham.

Vasiljevic, I., Chakrabarti, A., & Shakhnarovich, G. (2016) Examining the impact of blur on recognition by convolutional networks. arXiv preprint arXiv:1611.05760.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Wang, D., Forstmeier, W., Ihle, M., Khadraoui, M., Jerónimo, S., Martin, K., & Kempenaers, B. (2018) Irreproducible text-book “knowledge”: The effects of color bands on zebra finch fitness. *Evolution*, 72(4), 961-976.

de Weerd, N., van Langevelde, F., van Oeveren, H., Nolet, B. A., Kölzsch, A., Prins, H. H., & de Boer, W. F. (2015) Deriving animal behaviour from high-frequency GPS: tracking cows in open and forested habitat. *PloS one*, 10(6), e0129030.

Weinstein, B. G. (2018) A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3), 533-545.

Table 1. Rate of positive identification when testing in all contexts for the sociable weavers. Right column gives the identification success rate when noise and blurs were artificially added to training images to match the quality of testing images (see section "Testing models").

Perspective	Positive identification	Positive identification after adding blur and noise
Validation	0.924	0.903
Feeder (close)	0.926	0.926
Feeder (far)	0.958	0.972
Plate (ground)	0.867	0.944
Plate (top)	0.675	0.916