Mammalian Biology (2022) 102:875–897
https://doi.org/10.1007/s42991-022-00224-8

AUTOMATED INDIVIDUAL RECOGNITION

Pre-trained models are not enough: active and lifelong learning is important for long-term visual monitoring of mammals in biodiversity research—Individual identification and attribute prediction with image features from deep neural networks and decoupled decision models applied to elephants and great apes

Paul Bodesheim[1] · Jan Blunk[1] · Matthias Körschens[1] · Clemens-Alexander Brust[2] · Christoph Käding[2] · Joachim Denzler[1,2]

Received: 19 September 2021 / Accepted: 31 December 2021 / Published online: 12 April 2022

Abstract
Animal re-identification based on image data, either recorded manually by photographers or automatically with camera traps, is an important task for ecological studies about biodiversity and conservation that can be highly automatized with algorithms from computer vision and machine learning. However, fixed identification models only trained with standard datasets before their application will quickly reach their limits, especially for long-term monitoring with changing environmental conditions, varying visual appearances of individuals over time that differ a lot from those in the training data, and new occurring individuals that have not been observed before. Hence, we believe that active learning with human-in-the-loop and continuous lifelong learning is important to tackle these challenges and to obtain high-performance recognition systems when dealing with huge amounts of additional data that become available during the application. Our general approach with image features from deep neural networks and decoupled decision models can be applied to many different mammalian species and is perfectly suited for continuous improvements of the recognition systems via lifelong learning. In our identification experiments, we consider four different taxa, namely two elephant species: African forest elephants and Asian elephants, as well as two species of great apes: gorillas and chimpanzees. Going beyond classical re-identification, our decoupled approach can also be used for predicting attributes of individuals such as gender or age using classification or regression methods. Although applicable for small datasets of individuals as well, we argue that even better recognition performance will be achieved by improving decision models gradually via lifelong learning to exploit huge datasets and continuous recordings from long-term applications. We highlight that algorithms for deploying lifelong learning in real observational studies exist and are ready for use. Hence, lifelong learning might become a valuable concept that supports practitioners when analyzing large-scale image data during long-term monitoring of mammals.

Keywords Active learning · Animal re-identification · Attribute prediction · Continuous learning · Deep learning · Human-in-the-loop · Lifelong learning · Neural networks

Handling editors: Leszek Karczmarski and Stephen C.Y. Chan.

This article is a contribution to the special issue on "Individual Identification and Photographic Techniques in Mammalian Ecological and Behavioural Research – Part 1: Methods and Concepts" — Editors: Leszek Karczmarski, Stephen C.Y. Chan, Daniel I. Rubenstein, Scott Y.S. Chui and Elissa Z. Cameron.

✉ Paul Bodesheim
paul.bodesheim@uni-jena.de

[1] Computer Vision Group, Friedrich Schiller University Jena, 07737 Jena, Germany

[2] DLR Institute of Data Science, Mälzerstr. 3, 07745 Jena, Germany

Introduction

Many ecological studies for monitoring biodiversity or analyzing animal behavior require the identification of individuals. Typically, a lot of images or even videos are collected either automatically by camera traps or manually by photographers. The huge amount of collected image data then needs to be evaluated to first perform a visual identification and afterwards a downstream task like recognizing activities of individuals, or counting individuals to estimate population sizes and to track changes in population sizes over time. Instead of manually investigating the large image collections, algorithms from computer vision and machine learning enable an automatic identification to support practitioners.

Our work focuses on a general approach of utilizing image features extracted by deep neural networks as abstract but high-level visual representations of individuals and we exploit different convolutional neural network architectures for this task. We demonstrate that based on these representations, individuals can be distinguished automatically using classifiers trained with extracted features from annotated reference images. Hence, in our approach, we decouple the feature extraction with deep neural networks from the decision model used to perform the classification, in contrast to an end-to-end learning of features and decision rules. To show the general applicability of our approach, we consider several taxa in our experiments: elephants (African and Asian) and great apes (gorillas and chimpanzees) shown in Fig. 1.

Moreover, besides the identification of individuals, we also show the usefulness of the extracted image features for predicting several attributes of the animals. These are, for example, discrete attributes such as the gender (binary classification: male vs. female) or the age group (multi-class classification: infant, juvenile, subadult, adult and elderly). However, also continuous attributes like the age of the individual can be estimated from the extracted image features using regression approaches.

One advantage of our decoupled approach that separates the feature extractor from the decision model is that we can easily exchange the final part of the processing pipeline and use either classification models or regression models for the prediction, depending on the task that should be solved. We do not need to change or re-train the neural network that is used for feature extraction when switching to another task and can, therefore, exploit the richness and compactness of the feature representations for multiple prediction problems. Hence, pre-trained neural networks can be leveraged to avoid costly network training from scratch, which is often difficult with limited amount of labeled training data for an individual identification task, especially at the beginning of a monitoring study. In contrast, extracting features once and then training a conventional classifier, e.g., a linear support

Fig. 1 Overview of the different taxa and example images for the species that are considered for automatic visual identification in this paper: the African forest elephant (*Loxodonta cyclotis*) in the top left image (from the ELPephants dataset presented by Körschens and Denzler 2019), the Asian elephant (*Elephas maximus*) in the top right image (from a video provided by the CCC lab: https://ccconservation.org/), the Western lowland gorilla (*Gorilla gorilla gorilla*) in the bottom left image (from the dataset of Brust et al. 2017), and the chimpanzee (*Pan troglodytes*) in the bottom right image (from the C-Tai dataset of Freytag et al. 2016)

vector machine (SVM) or a Gaussian process (GP) model, can be done quickly and in the field.

A second advantage of our decoupled approach is the easy integration of lifelong learning (Käding et al. 2016d). Instead of learning identification systems only once before the application using a fixed training set of annotated images, we also want to take changes into account, which naturally arise during the application when monitoring animals for a longer time period. For example, new individuals might enter the scenes that have not been recorded before and, therefore, need special treatment. Even known individuals might occur in unusual poses, their visual appearances change due to aging or injuries, or they are partially occluded in the images by other animals or vegetation. To avoid false predictions of an automatic system in such challenging scenarios, feedback from experts should be incorporated to resolve these hard cases where the algorithms have problems or are most uncertain. Furthermore, not all incoming images need to be checked manually, and instead, the recognition system could automatically select only a subset of most relevant images whose labels are then verified by the experts to reduce the annotation efforts. This selection process is the main task of active learning (Settles 2009), a key ingredient of lifelong learning to incorporate expert knowledge and feedback via a concept called human-in-the-loop (Käding et al. 2016d).

We, therefore, propose the combination of our decoupled approach with lifelong learning to incorporate additional knowledge over time by requesting further annotations for selected images during the application (Käding et al. 2016d). This allows for continuous improvements of the automatic identification system by exploiting newly recorded data and corresponding labels provided by experts. These labeled images are used to update the decision model of our approach (a classifier or a regression model) with incremental learning techniques to avoid costly re-training from scratch. This leads to a feedback loop of continuously requesting further annotations of domain experts to improve the recognition system over time, and this feedback loop can be repeated many times. Hence, learning to distinguish individuals can be improved over and over again by incorporating more and more annotated images during the application.

With lifelong learning, it is meant that the system learns during its whole life span that is basically defined by the duration of its application. This is in contrast to the common approach of taking any pre-trained neural network with parameters learned on standard datasets, probably slightly adapting it for the target task by fine-tuning on a small target dataset, and trying to solve the identification task with this fixed recognition model. While such an approach is nowadays easily achievable due to many well-documented deep learning frameworks that provide different neural network architectures with pre-trained parameters, its rating might be questionable. Depending on the study design and the benchmark used for evaluation, fixed pre-trained models can achieve good performance on rather small-scale datasets when there are only little variations of the underlying data distribution. However, long-term monitoring in practice has to cope with several additional challenges as mentioned above, such that the recognition system needs to adapt to new situations and circumstances, ideally by also incorporating expert feedback to verify certain cases that are difficult to decide. Those mechanisms are provided by lifelong learning, and we argue that considering long-term monitoring of mammals as an application for lifelong learning leads to continuous model improvements and increased recognition performance.

Our experimental results in "Experiments" are a collection of different studies (Freytag et al. 2016; Käding et al. 2016a; Brust et al. 2017; Körschens et al. 2018; Körschens and Denzler 2019) that we conducted during the last years in the context of animal re-identification and attribute prediction, now put under one umbrella because the developed recognition systems for different animal species share the same general approach in terms of algorithmic design. While the original work published at computer vision and machine learning venues aimed at describing the developments of the different algorithms from a technical perspective, we consider them here from the application point of view. We put them in relation to each other in context of the same general approach, and with their basic ability to incorporate lifelong learning techniques that can directly be applied to the incorporated decision models for improving them. We also present new results for identifying Asian elephants in camera trap videos ("Identifying Asian elephants") and highlight the advanced possibilities for monitoring animals beyond identifying individuals. These possibilities are on the one hand attribute predictions of individuals and on the other hand lifelong learning with human-in-the-loop because the latter becomes more and more important for real applications and long-term monitoring (Stewart et al. 2021). Lifelong learning and the involved active sample selection via active learning strategies helps to reduce the workload of trained field experts for annotating data, since it becomes impossible to manually inspect all images due to the huge amount of recordings that are collected over time.

Related work

In this section, we briefly review related work on algorithms for the tasks we want to solve. After mentioning relevant work on image classification and object detection in general ("Image classification and object detection"), we also list approaches for fine-grained recognition of different animal species ("Fine-grained recognition and species

classification"), because it is a highly related task similar to identifying individuals ("Identifying individuals") from a machine learning perspective. Furthermore, we discuss previous work on lifelong learning with a particular focus on active learning and human-in-the-loop aspects ("Lifelong learning and active learning").

Image classification and object detection

Image classification methods can be used to assign discrete labels such as the ID of an individual animal to an entire image. In case of multiple animals in a single image, it makes sense to localize each individual in an image, e.g., by a rectangular bounding box, and assign an ID to each animal separately. This joint estimation of bounding boxes and class labels (IDs) is in general called object detection. In the following, we review related work on image classification and object detection based on recent deep neural network developments.

Deep neural networks for image classification

Deep neural networks and in particular convolutional neural networks (CNNs) have already been developed decades ago (LeCun et al. 1989; Matan et al. 1990; LeCun et al. 1990; LeCun and Bengio 1995; LeCun et al. 1998), and they consist of multiple data processing units arranged in separate but interconnected layers to transform input data like images directly into the desired outputs like classification scores used for individual identification. However, their major breakthrough in the computer vision community and especially for the task of image classification has been achieved by Krizhevsky et al. (2012). With the support of powerful graphical processing units (GPUs), they have been able to train a large CNN architecture often referred to as AlexNet on the well-known ImageNet dataset (Deng et al. 2009; Russakovsky et al. 2015) and improved classification accuracy by a large margin compared to competing approaches. Since then, various neural network layers and different architectures have been proposed that further improved the performance for automatic image classification, for example, ResNet (He et al. 2016) and Inception (Szegedy et al. 2016) architectures besides others (Simonyan and Zisserman 2015; Chollet 2017; Xie et al. 2017). In particular, the size of the networks has grown over time, with an increasing number of layers and network parameters to enable more expressive power of the learned feature representations. Although developed for object recognition in general, these deep neural networks are able to learn rather generic feature representations from example data. Hence, neural networks pre-trained on object category datasets like ImageNet (Russakovsky et al. 2015) can also be applied as feature extractors to obtain reasonable numerical representations for images of objects in a specific domain, e.g., for monitoring certain animals. Sometimes, the networks are slightly adapted to the new domain by a transfer learning technique called fine-tuning (Yosinski et al. 2014; Sharif Razavian et al. 2014; Long et al. 2019). Transfer learning means that the knowledge extracted from one dataset during the training of a neural network for a certain classification task can be transferred to and exploited for another classification task, e.g., using a neural network for identifying individuals that has originally been trained to recognize general object categories. Furthermore, CNNs often serve as backbone network architectures for object detection methods like the ones listed in the following.

Object detection with deep neural networks

There exist many approaches for localizing objects in images, however, in the following, we focus on those based on deep learning methods. One of the first deep object detectors has been proposed by Girshick et al. (2014) and is called region-based CNN (R-CNN). Object proposals determined with an unsupervised method are selected as candidates, for which features are extracted by a backbone CNN and classified by a support vector machine (SVM), one for each class, to determine the corresponding class label. This strategy improved previous sliding window-based approaches for object detection (Dalal and Triggs 2005; Felzenszwalb et al. 2010) and our approach for identifying individuals described in "Methods" follows a similar concept. Note that the R-CNN approach has been extended and improved in several ways (Girshick 2015; Ren et al. 2015; He et al. 2017).

Another general object detection approach based on deep learning is YOLO proposed by Redmon et al. (2016). This detector is trained end-to-end and is able to obtain all detections for an image with a single forward pass of the network and minimal post-processing operations. Another state-of-the-art approach for object detection is single-shot detection (SSD) proposed by Liu et al. (2016). Similar to YOLO, they also use only a single forward pass per image but use a more complex output encoding together with assumptions about the aspect ratios of bounding boxes as well as predictions on different scales. Further improvements of the YOLO approach have been published as YOLOv2 (Redmon and Farhadi 2017), like more fine-grained feature maps and the awareness of multiple object scales by resizing the network during training. In the context of wildlife monitoring, dedicated object detectors based on R-CNN or YOLO have been trained (Parham et al. 2018; Beery et al. 2019; Tabak et al. 2019).

Fine-grained recognition and species classification

Fine-grained recognition denotes an image classification task for which the classes often differ only slightly in small details because objects or instances belong to the same domain. This domain is typically given by a joint superclass, e.g., identifying different car models where all objects are cars (Krause et al. 2013). However, fine-grained recognition is more frequently applied in the context of animal species classification, e.g., distinguishing different bird species (Wah et al. 2011; Cui et al. 2018; Korsch et al. 2019, 2021b) or moth species (Rodner et al. 2015; Böhlke et al. 2021b, a; Korsch et al. 2021a). Besides early studies on fine-grained recognition with deep neural networks (Branson et al. 2014; Rodner et al. 2016), many approaches have been developed that can coarsely be partitioned in two subsets.

The first subset contains global approaches that solely process the entire image with a neural network as in standard image classification, either relying on smart pre-training and transfer learning (Krause et al. 2016; Cui et al. 2018) or applying advanced feature pooling strategies for aggregating localized visual information (Lin et al. 2015; Gao et al. 2016; Simon et al. 2020). On the other hand, the second subset denotes part-based and attention-based approaches (Ge et al. 2019; He et al. 2019; Korsch et al. 2019; Zhang et al. 2019), which rely on detecting relevant image regions often associated with semantic parts such as the beak, the belly, and the wings of a bird. These regions are then encoded using feature representations from deep neural networks that should help focus on the small details for distinguishing visually similar classes. While automatic species classification is already quite challenging due to the high visual similarity of different species from the same class, order, or family, the distinction of individuals from the same species is more complicated since relevant features might be even harder to determine.

Note that species recognition is not always a fine-grained recognition problem, e.g., when considering camera traps in forests or national parks where the system needs to distinguish different wild animals such as deer, fox, and wild boar. However, these species recognition tasks are also commonly tackled with deep neural networks (Villa et al. 2017; Norouzzadeh et al. 2018; Willi et al. 2019; Tabak et al. 2019; Schneider et al. 2020a).

Identifying individuals

Traditional methods for identifying individuals of a certain animal species follow a non-invasive genetic mark-recapture approach that allows for precise estimates but requires high levels of expertise leading to limited scalability (Kühl 2008; Guschanski et al. 2009; Arandjelovic et al. 2010; Roy et al. 2014). Camera traps offer a cheap and widely accessible alternative for long-term usage (Schneider et al. 2019), e.g., in combination with distance sampling (Howe et al. 2017) or capture-recapture models (Kühl 2008; Pebsworth and LaFleur 2014). Thus, recording large amounts of visual data for monitoring purposes also requires computer vision algorithms for automatic evaluations (Schneider et al. 2019), since manual investigations would be too time-consuming (Schneider et al. 2020a).

The field of animal biometrics (Kühl and Burghardt 2013) is dedicated to detecting visual patterns that enable the distinction of individuals (Crall et al. 2013; Cheema and Anand 2017) and different systems have been developed for animal detection and re-identification (Crall et al. 2013; Berger-Wolf et al. 2017; Parham et al. 2018; Yang et al. 2019; Bakliwal and Ravela 2020). However, recent advances in recognizing human faces (Taigman et al. 2014; Schroff et al. 2015; Parkhi et al. 2015) have inspired the identification of great apes (Loos et al. 2011; Loos 2012; Loos and Ernst 2013; Brust et al. 2017; Freytag et al. 2016; Crunchant et al. 2017; Schneider et al. 2020b) and elephants (Ardovini et al. 2008; Körschens et al. 2018; Körschens and Denzler 2019; Kulits et al. 2021) based on the detected faces of the animals. Further animals that have been considered for automatic re-identification are tigers (Shukla et al. 2019; Yu et al. 2019; Liu et al. 2019; Weideman et al. 2020) and turtles (Carter et al. 2014; Dunbar et al. 2021), as well as ringed seals (Nepovinnykh et al. 2020) and manta rays (Moskvyak et al. 2020).

Lifelong learning and active learning

A typical machine learning workflow in an academic setting, including a part of the experiments in this work, is best described in terms of a waterfall model (Data Science Process Alliance 2021). The discrete steps of collecting data, deriving and optimizing a model, and finally deploying it, are performed in order and once each. In contrast, lifelong learning (Käding et al. 2016d) assumes a continuous stream of unlabeled data. Repeatedly, a small fraction is selected for labeling by active learning methods (Settles 2009; Freytag et al. 2014; Käding et al. 2016a, 2018; Wang et al. 2017; Brust et al. 2019) and then used to update the model incrementally (Käding et al. 2016c; Rebuffi et al. 2017; Castro et al. 2018). The continuity of lifelong learning aligns well with long-term monitoring projects, and active learning specifically is identified as a promising research avenue in this context (Norouzzadeh et al. 2018). Drawbacks of the waterfall learning approach for monitoring are illustrated by Beery et al. (2018) with solutions involving repeated training proposed by Beery et al. (2019).

Active learning improves annotation time efficiency in animal presence detection (Käding et al. 2016a), species classification (Evans et al. 2014; Brust et al. 2020) as well

as re-identification tasks (Norouzzadeh et al. 2020). Continuous data and model updates are posed as a fundamental problem for animal identification by Stewart et al. (2021). An annotation interface for ecological monitoring studies is provided by Kellenberger et al. (2020) and a graphical user interface implementing a lifelong learning approach for animal monitoring is described by Brust et al. (2021).

Methods

In this section, we first describe our general approach of using image features from deep neural networks and decoupled decision models to perform the prediction task ("General approach using deep image features and decoupled decision models"). We then characterize the two involved building blocks individually: the feature extraction using a backbone CNN network that is restricted to the image region of the localized animal and its head ("Individual localization and deep image feature extraction"), and the application of standard machine learning models for classification or regression to perform an identification task or to predict attributes of the animals ("Classification and regression for individual identification and attribute prediction"). Decoupling the feature extraction from the final decision model for predicting the desired outputs easily allows for incorporating lifelong learning, where we want to integrate further knowledge via expert annotations for new images following the human-in-the-loop concept ("Lifelong learning with human-in-the-loop"). However, integrating human feedback typically requires active learning strategies for selecting the most relevant images that need to be annotated to gain most from the additional annotation efforts, and incremental learning techniques to update the models ("Active learning strategies and incremental learning").

General approach using deep image features and decoupled decision models

Our general approach is visualized in Fig. 2 and the main idea is to decouple the feature extraction with powerful CNN architectures from the final decision model that performs the prediction based on the extracted feature representations. To obtain meaningful features for an identification task, the feature extraction is restricted to the image region covered by the individual or even more localized to the corresponding head region, which is obtained by an object detection approach. Afterwards, the extracted features can be used to perform either individual identification or attribute prediction with dedicated decision models, which are machine learning models either for classification to obtain discrete outputs or for regression to obtain continuous outputs. In the next section, we specifically focus on the feature extraction for individuals which first need to be localized in the image. Note that localizing an individual or its head can be skipped if one is interested in an identification of individuals in manually taken photographs, where each image already contains only a single individual or its head in a close-up view. However, for a wider range of applications including the identification in images from camera traps, we also include the localization step.

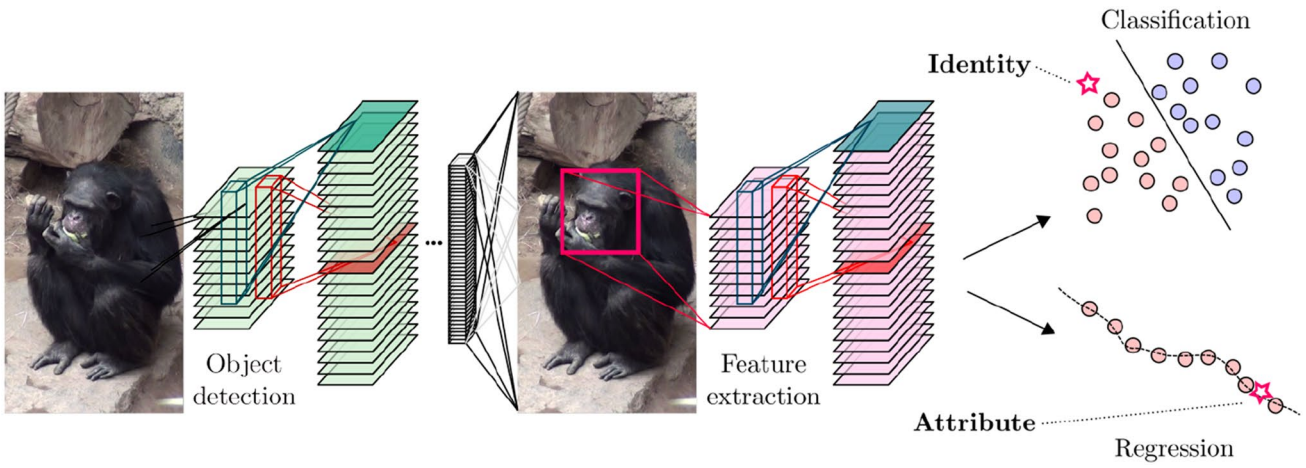

Fig. 2 Overview of our general approach: the head of an individual is first detected with a deep learning object detector, then features from the corresponding image patch are extracted using a pre-trained deep neural network architecture. Finally, the extracted features can be used by a decision model to predict the desired outputs, e.g., either a classifier to obtain the corresponding ID or a regression method to obtain continuous attribute values. The chimpanzee image is taken from the C-Zoo dataset of Freytag et al. (2016)

Individual localization and deep image feature extraction

Given an image containing one or multiple animals of the same species, we first run an object detector to localize each individual and determine a corresponding bounding box. For this task, a deep learning detector such as R-CNN or YOLO can be used as mentioned in "Image classification and object detection". Since one usually has not enough training data for a single species to train such a detector from scratch, pre-trained detection networks for standard object categories are used and fine-tuned with a small set of annotated images with animals from the target species.

As confirmed by domain experts, the head of an animal usually contains many discriminative features for distinguishing individuals, and it, therefore, makes sense to focus on the head regions for extracting meaningful feature representations. Hence, one can directly fine-tune the detector only for the heads and not for the whole bodies, which requires corresponding bounding box annotations. In case of great apes, it is then advantageous to fine-tune a face detector trained to localize human faces in images instead of a detector for common objects, mainly because of the higher similarity of ape faces to human faces and because of the large amount of available datasets for human face recognition that can be used to pre-train the detector (Schroff et al. 2015; Parkhi et al. 2015; Taigman et al. 2014).

Given the corresponding bounding boxes, we extract feature representations from these image regions by applying deep neural networks to the associated image patches. Due to the limited amount of annotated data that is usually available for an identification task, it is not possible to learn parameters of large, powerful neural network architectures only from the images of the species under investigation. The reason for this is that there are way more parameters to estimate for large network architectures compared to the number of training images, which leads to heavy overfitting of the learned model to the training dataset because the model would be able to memorize each individual example. In the supplemental material of Käding et al. (2016c), we show that the negative impact of fine-tuning only the last layer, as opposed to more layers or the whole network, is negligible even when there is sufficient training data. To leverage their advantages in our restricted training data scenarios, we apply pre-trained neural networks as black-box feature extractors, where the network parameters have been determined using large-scale datasets of common object categories, e.g., ImageNet (Russakovsky et al. 2015). Common architectures like AlexNet (Krizhevsky et al. 2012) or residual networks like ResNet50 (He et al. 2016) are implemented in various deep learning frameworks and can easily be applied off-the-shelf. Given the extracted feature representations, a classifier can be learned from annotated training data using these features as inputs and delivering a class label that is associated with the ID of an individual. However, classifiers can also be used for predicting discrete attributes from the extracted feature representations, and methods for regression allow the estimation of continuous attributes like the age of an individual.

Classification and regression for individual identification and attribute prediction

Since we decoupled the feature extraction with deep neural networks from the final prediction for the corresponding task, we are free to choose another machine learning model that operates on the extracted features for estimating animal IDs or attributes. This final model for prediction can be substituted according to the task that needs to be solved, and it could either be a multi-class classifier for animal re-identification or age group prediction, a binary classifier for gender prediction, or a regression model for estimating continuous values like the age of an animal.

For classification tasks, we rely on support vector machines (SVMs) that have been widely used for both binary and multi-class classification (Cortes and Vapnik 1995; Schölkopf and Smola 2001). The idea of SVMs is to find decision boundaries in the feature space that separate different classes (e.g., individual animals) with a maximum margin, i.e., maximizing the distance between the examples of each class and the decision boundary while ensuring that examples of the same class are on the same side of the decision boundary. With SVMs it is possible to determine linear decision boundaries, i.e., hyperplanes in the feature space via linear SVMs, or non-linear decision boundaries using kernel functions and kernel SVM (Schölkopf and Smola 2001).

Although there exist support vector approaches for regression (Schölkopf and Smola 2001), we use Gaussian processes (GPs) for regression tasks (Rasmussen and Williams 2006). They have the advantage that uncertainties can be estimated together with the predicted output, because the model is formulated in a probabilistic manner, and that closed-form solutions exist for learning the model parameters under the assumption of a Gaussian noise model. Note that GPs can also be used for classification via label regression (Rasmussen and Williams 2006; Kapoor et al. 2010; Rodner et al. 2017). Furthermore, SVMs and GPs have the advantage that there exist update rules and algorithms for incremental learning and incorporating additional data (Cauwenberghs and Poggio 2001; Diehl and Cauwenberghs 2003; Lütz et al. 2013; Freytag et al. 2014), which is an important aspect for lifelong learning.

Alternatives such as random forests, nearest-neighbor rules, and other classifiers are not considered in our evaluation. Since the features generated by a pre-trained neural

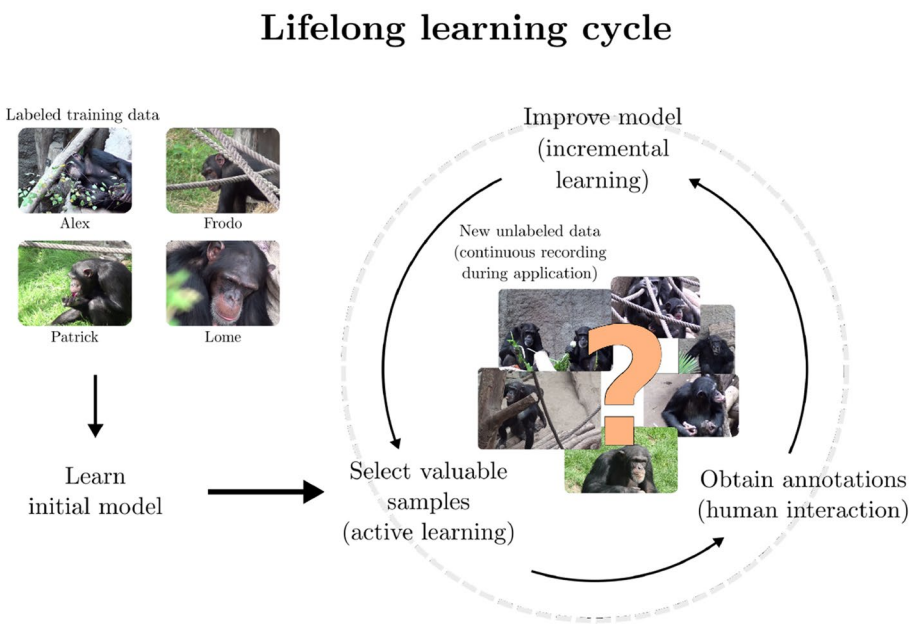

Fig. 3 Overview of our proposed lifelong learning cycle realizing the human-in-the-loop concept. After learning an initial model with annotated training data, a continuous feedback loop is established that leverages additional data recorded during the application. Active learning strategies select valuable examples from the incoming data stream that are then annotated by experts (human interaction) and finally used to update and improve the model via incremental learning. Chimpanzee images are taken from the C-Zoo dataset of Freytag et al. (2016)

network are trained to be linearly separable by definition, there is little need for highly non-linear or overparameterized classifiers. Instead, we leverage the high computational efficiency of SVM and GP models.

Lifelong learning with human-in-the-loop

Lifelong learning denotes a concept, where learning a recognition system is not fully automatic, but where human knowledge and intervention is integrated in the training algorithms to improve learning and to make the final decisions of the recognition system more robust. It can be seen as a semi-automatic learning approach and to emphasize the human interaction part, it is called learning with human-in-the-loop. The loop is usually referred to a continuous training process of a recognition system in the context of long-term applications where additional data becomes available over time. We visualize this loop in a lifelong learning cycle in Fig. 3. Given an initial model learned with labeled training data, the following steps are repeated over and over again during an application, where new unlabeled data becomes available due to continuous recordings. First, active learning strategies are applied to select valuable samples that are supposed to be the most interesting and important ones with potentially the largest impact on improving the current model. Of course the main question here is how to select these samples and details are given in the next section ("Active learning strategies and incremental learning"). The selected samples are then passed to the domain experts for annotation. This is the human interaction part and one common goal is to reduce the labeling efforts of the experts or to keep it on a reasonable and maintainable level because manual inspection is usually a time-consuming task. Finally, we have additional labeled data provided by expert feedback, which can be used to improve the model via incremental learning and corresponding model parameter updates. In the next step, the loop starts again with automatically selecting samples from the incoming data. Due to incorporating expert feedback during learning, the continuous training loop is also called feedback loop.

Note that one can further think of a combined recognition approach, where the automatic system makes several proposals for an individual and an expert has to decide for the final ID label based on the provided set of candidates. In this way, the system supports the experts in the application by reducing the set of possible choices, and we reflect such an option in our experimental results by providing also top-5 accuracies, meaning that the system is successful when the correct individual is among the five highest ranked IDs, besides top-1 accuracies, which are used for evaluating a fully automatic identification. However, in a scenario where the system makes several suggestions per individual, an expert is involved in every decision for every recording, which might be helpful for studies at smaller scales but becomes infeasible for large-scale monitoring activities.

We, therefore, consider human-in-the-loop for improving the training of the recognition system over time in a lifelong learning setup, where we get access to additional image data during the application. To learn from the additional data, appropriate annotations for the new data are required. While it would also be possible to request bounding box annotations, which could even be provided by laymen who do not need to be able to distinguish individuals, we rely on the general detectors mentioned in "Individual

localization and deep image feature extraction" to localize animals in the images automatically and focus on requesting ID labels or attribute annotations from experts. Note that few additional bounding box annotations might be helpful for camera trap videos, in which multiple individuals need to be tracked and identified (see "Identifying Asian elephants" for a corresponding application). In this case, annotating the first few frames with locations and IDs could be helpful to improve the final performance of the system.

However, when we want to integrate additional image data in the learning process, the main challenge is to select which images need to be annotated by experts to gain the largest improvements of the recognition system because it is usually too expensive to label all incoming data. On the one hand, it is too expensive in terms of human resources and time that is spent by trained experts for the labeling. They might not even be able to go through all the recorded images in reasonable time, since the datasets of large-scale long-term monitoring studies are simply overwhelming and continuously increasing. On the other hand, it is too expensive from a computational resources point of view. There is only limited data storage available such that not all incoming data can be stored permanently. Furthermore, training or updating a model with an ever increasing size of the training dataset takes large computational time, also with sophisticated incremental learning techniques, and might even become infeasible at some point if there is no appropriate sample selection involved.

Therefore, useful data selection strategies have been developed as the key ingredient for active learning methods (Settles 2009; Freytag et al. 2014; Käding et al. 2016a, 2018; Wang et al. 2017; Brust et al. 2019) to tackle the sample selection problem. One obvious requirement for an application in unconstrained environments is the annotation of data that belongs to a new and previously unknown individual, i.e., which has not been observed in the training dataset. Updating the classifier with data of new individuals allows for increasing the number of individuals the system can recognize over time. While there exist methods that specifically focus on automatically detecting instances of unknown classes known as novelty detection (Bodesheim et al. 2013, 2015), active learning strategies also incorporate mechanisms for choosing images of unknown individuals that should be presented to an expert for annotation (Käding et al. 2015). Furthermore, related work on open set recognition (Scheirer et al. 2013, 2014) and open world recognition (Bendale and Boult 2015) deal with classification problems, for which the number of different classes is not fixed but might grow during the application.

For performing active learning with human-in-the-loop in a lifelong learning application such as animal monitoring and identification, a continuous feedback loop needs to be implemented that consists of several steps: selecting a subset of the incoming data for annotation, requesting the labels from experts, incorporating the additional data in the learning process of the system, and updating the model to enable improved predictive performance. We use the WALI framework (Käding et al. 2016d) for this purpose, where WALI stands for watch-ask-learn-improve and resembles the four steps mentioned before. Note that these steps can be repeated many times during an application that continuously delivers new data, which leads to a so-called lifelong learning cycle (Käding et al. 2016d). After improving the model with learned parameter updates from additionally annotated data, the incoming data stream is further monitored and analyzed (watch), selecting the presumably most relevant images for requesting expert annotations (ask), and learning from these additional data samples (learn) to update the model parameters and obtain an improved recognition system (improve). This feedback loop ensures that the most current knowledge influences the selection performed by active learning. For more details about WALI, we refer to the original work of Käding et al. (2016d).

To summarize, the two most important aspects in a lifelong learning cycle are a clever selection of data for annotating and efficient algorithms for incremental learning to update the model parameters continuously. Hence, we discuss active learning strategies and incremental learning techniques in the following.

Active learning strategies and incremental learning

To update a recognition model with additional training samples, interesting and relevant samples have to be selected from the incoming data during the application such that experts can annotate them. The goal of active learning is to choose those samples that improve the recognition system the most when they have been labeled and used to update the model parameters. There exist several strategies for selecting meaningful samples, and the most prominent and most intuitive one is uncertainty sampling (Lewis and Gale 1994; Settles 2009; Wang et al. 2017). This means that those images are selected for annotation, for which the current model is most uncertain about its prediction. Hence, a human expert should resolve such situations, e.g., if a classifier is less confident and favors multiple classes (IDs) for a single individual.

Typically, a classifier provides a soft output score, i.e., a continuous scalar value, for each class. This value reflects the confidence of the classifier that the corresponding class is present in the image under investigation. Sometimes, the output scores are probabilities for observing a certain class but stochastic properties (outputs are between zero and one, and sum up to one over all classes) are in general not

required, since simply the class with the largest output score can be assigned. Hence, the soft output score can in principle directly be used as a measure of confidence for the classifier. Another easy implementation of uncertainty sampling for a classification model is the best-vs-second-best approach also known as margin sampling (Settles 2009; Wang et al. 2017). If the difference between the highest score and the second-highest score of a classifier is small, the model is highly uncertain about assigning one of these two classes to an image. Formally, the feature representation lies close to the decision boundary of the classifier in the feature space, which increases the risk of a misclassification. In this case, an expert should provide the correct label.

However, extending models with samples selected by an uncertainty criterion does not always lead to the best model improvements in terms of achievable recognition performance, which we also show in our experiments in "Lifelong learning of age predictors". There exist more advanced active learning strategies, e.g., selecting unlabeled samples based on the expected model output change (EMOC) criterion (Freytag et al. 2014; Käding et al. 2016a, b, 2018). With EMOC, those samples are selected for annotation, which are expected to change the future model outputs the most after they have been used to update the model parameters. This makes sense because highly informative examples are likely to cause large changes of model outputs for future predictions after the model updates, and therefore, those examples are most valuable for the model evolution. Hence, samples selected by the EMOC criterion have a large impact on the behavior of the model, and it has been shown that EMOC is able to select the most influential examples including those of unseen categories or objects in new poses for classification tasks (Freytag et al. 2014; Käding et al. 2016a, b).

Another advantage of EMOC is that it can not only be used for classification, but also for regression models (Käding et al. 2018). Although with Gaussian processes for regression, one can directly compute the uncertainty of an estimate in terms of the predictive variance (Rasmussen and Williams 2006), it has been shown empirically that EMOC selects samples that are more beneficial for faster learning of accurate models (Freytag et al. 2014; Käding et al. 2016a, b). Hence, we propose using the EMOC criterion to select samples for any task, no matter if it is a classification task (individual identification or discrete attribute prediction) or a regression task (continuous attribute prediction).

Once the additional samples have been annotated by domain experts, the recognition model needs to be updated via incremental learning techniques. Although there exist approaches for updating deep neural networks in continuous learning via fine-tuning (Käding et al. 2016c), there are several challenges that arise in a lifelong learning scenario (Lomonaco and Maltoni 2017; Maltoni and Lomonaco 2019), such as catastrophic forgetting (McCloskey and Cohen 1989; Robins 1993; French 1999; Shmelkov et al. 2017; Hayes et al. 2020). Since in our general approach (see "General approach using deep image features and decoupled decision models"), pre-trained neural networks are used for feature extraction only because feature extraction is decoupled from solving the final downstream task with a classification or regression model, we do not update neural network parameters and keep the feature extraction network constant. Note that this is even in line with attempts for end-to-end incremental learning of neural networks (Castro et al. 2018), where the feature extraction part is kept unchanged as well. Hence, we also only update the final classification or regression models, which enables a tight feedback loop between annotators and active learning methods due to very quick model updates. For SVM classifiers, there are efficient update algorithms for incremental learning available (Cauwenberghs and Poggio 2001; Diehl and Cauwenberghs 2003), and in case of Gaussian process regression models, there exist closed-form solutions for incremental learning (Lütz et al. 2013; Freytag et al. 2014).

Experiments

With our experiments, we demonstrate that our general approach works well in numerous applications, including different animal species that are considered as well as different tasks that need to be solved (identification and attribute predictions). We first show results for identifying individual elephants ("Identification of African forest elephants" and "Identifying Asian elephants"), followed by the identification of gorillas ("Identifying gorillas") and chimpanzees ("Identifying chimpanzees"). Then, our results for predicting attributes of animals are presented ("Attribute predictions"). Finally, we demonstrate the application of lifelong learning, for which we picked the task of age prediction ("Lifelong learning of age predictors"). Note that all datasets involved in our experiments including the one for lifelong learning are of rather small scale in terms of number of example images. Nevertheless, the benefits of lifelong learning are already visible and we expect even further improvements of the recognition systems in actual long-term monitoring applications.

Identification of African forest elephants

The first experiment we present is about the identification of African forest elephants (*Loxodonta cyclotis*) (Körschens and Denzler 2019), using images that are recorded manually by photographers. We start with a short description of the underlying dataset.

ELPephants dataset

For identifying African forest elephants, we use the ELPephants dataset presented by Körschens and Denzler (2019), which contains images from a long-term monitoring study of elephants in the Dzanga bai clearing of the Dzanga-Ndoki National Park in the Central African Republic. The dataset covers 276 elephant individuals and consists of 2078 manually taken images taken over about 15 years, with each image having a single annotation for the depicted individual. Although there is little to no occlusion of the animals in the images due to manually taken photographs in the clearing, there are several other challenges besides varying viewpoints, like occlusion of useful features for identification through mud, aging of individuals during this long time span, as well as changing appearance of individuals due to new scars and broken tusks due to fights that might occur over time. For more details about this dataset, we refer to the descriptions of Körschens and Denzler (2019). In our experiments, the dataset is split randomly into 1573 images for training and 505 images for testing using a 75%/25% stratified split. Since our approach to this specific problem assumes a closed set of individuals, the number of individuals in each split is the same and there are no held-out individuals in the test set.

The elephant identification system (EIS)

For the identification of individual elephants, there are several characteristic features such as size and shape of the tusks, which also vary between male and female elephants (Körschens and Denzler 2019). Furthermore, signs from fights or other injuries are important, like broken tusks as well as scars, rips or holes in the ears. We, therefore, focus on the head including ears, tusks, and trunk of the elephants for the identification. Our elephant identification system (EIS) (Körschens and Denzler 2019) uses an elephant head detector, which is a YOLO network (Redmon et al. 2016) trained on 1285 elephant images from Flickr.[1] These images are not part of the dataset for identification and have been manually annotated with bounding boxes covering head, ears, tusks, and trunk of each individual (Körschens and Denzler 2019). When applying the head detector on the identification dataset for which we only have a single ID label per image, we might obtain multiple bounding boxes due to noise or multiple elephants present in the image, and we need to keep only one bounding box for the identification. Hence, we select the most prominent bounding box as the one that covers the largest image area, weighted by the confidence score of the detector (Körschens and Denzler 2019).

[1] https://www.flickr.com/.

For the selected head region, features are extracted using a ResNet50 network (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015), and we compare different network layers with respect to their representation power. Feature extraction and model training was performed on the original images and their horizontally flipped versions to account for an appropriate data augmentation strategy. We then use a linear SVM classifier to perform the identification task. To enable the possibility for considering multiple images of the same individual for the identification task in an extended application, e.g., when there are short videos available or when the photographer takes multiple images of the same individual on the same day (perhaps from different viewpoints), a simple aggregation step can be added. In case of multiple images for a single decision, we follow the concept of late fusion and combine the classifier outputs of the individual images by averaging class confidence vectors obtained from the SVM classifier to obtain the final classification scores. This allows for easily integrating images of different viewpoints to allow for a more robust identification.

Elephant identification results

Our EIS described before consists of an elephant head detector, followed by feature extraction from the head region using a deep neural network, and classification of the head for identification. The head detector achieves an average precision of 90.78%, evaluated on 227 manually annotated test images (Körschens and Denzler 2019).

For identifying individuals, we have tested different activation layers from the ResNet50, whose outputs are used as a feature representation to describe an elephant head and to perform the classification. It turned out that the activation layers from the 13th and 14th convolutional block, in the following denoted by `activation40` and `activation43`, performed best. Hence, we only report results for these layers, for which we also added different maximum pooling layers to account for translation invariance.

The results (top-1 and top-5 accuracies) are shown in Table 1, and we observe that the best results are achieved with features from the `activation40` layer. When taking only a single image for the identification of an elephant into account, the best result of our EIS is a top-1 accuracy of 56.0% for fully automatic identification with a region size for maximum pooling of 6×6. Regarding the top-5 accuracy, pooling with a smaller region size (5×5) performed slightly better, leading to 72.6%. We also performed experiments where we used two images for a single decision and were able to improve recognition accuracies significantly. With two images, a top-1 accuracy of 74.2% is achieved with features from the `activation40` layer and maximum pooling with a region size of 6×6. Furthermore, the correct individual is among the top-5 suggestions of the system in

Table 1 Individual identification results for African forest elephants on the ELPephants dataset (Körschens and Denzler 2019)

Activation layer	Max pooling filter size	Results for one image per individual		Results for two images per individual	
		Top-1 accuracy (%)	Top-5 accuracy (%)	Top-1 accuracy (%)	Top-5 accuracy (%)
activation40	4×4	50.8	70.6	69.8	81.8
activation40	5×5	54.4	**72.6**	71.4	83.2
activation40	6×6	**56.0**	71.6	**74.2**	**85.2**
activation43	4×4	52.2	71.6	70.0	83.0
activation43	5×5	54.6	70.8	72.2	83.2
activation43	6×6	52.4	70.0	70.8	82.8
activation43	No pooling	51.8	65.9	68.6	80.4

Best results indicated in bold

85.2% of the cases, which would already narrow down the final decision from 276 possible individuals to support the ecologists. Hence, our experimental results show that if it is possible within the application, incorporating more than one image for the identification is beneficial and more robust because more visual information due to different viewpoints can be exploited.

Note that the identification of individual elephants based on whole bodies compared to considering only head regions has led to worse performance, and the same holds for taking the last activation layer of the ResNet50 to extract meaningful features (Körschens and Denzler 2019), which is typically done when using a deep neural network as a feature extractor. Hence, in our experiments we have shown that earlier layers within the ResNet50 carry more semantically meaningful information for identifying individual elephants.

The intended application for this specific approach assumes a closed set of individuals. Furthermore, the predictions are only used as suggestions to assist human annotators, who still manually assess each image, but can do so faster with the help of our system. Because of the small number of images and the high variance in quality, a fully automated approach is not feasible and human interaction is required. Hence, false predictions or out-of-dataset individuals are not particularly harmful in this application and only impact efficiency.

Identifying Asian elephants

In this section, we demonstrate the application of our EIS from "The elephant identification system (EIS)" for identifying Asian elephants (*Elephas maximus*) in short video clips. First, the dataset for our experiment is described.

Asian elephant video dataset

The dataset has been provided by a research group from the comparative cognition for conservation laboratory (CCC lab[2]) at Hunter College, City University of New York. It contains annotated camera trap footage of Asian elephants recorded automatically in forest areas in Thailand. There are 108 individuals spread over 683 short videos, each having a length of roughly 20 s. Note that there is often more than one individual present in a video and ideally all of them should be identified in each frame. Interestingly, 274 videos are recorded during nighttime, posing an additional challenge for the identification of individuals. A further challenge is imposed by younger elephants, since 26 of the 108 individuals are offsprings, which often have less distinctive features compared to adult elephants. For evaluation, the dataset is split into 508 videos for training and 175 videos for testing.

Pipeline for processing video data

To exploit the opportunity of having video data for identification, we want to use as many frames within a video as possible that contain a single individual to integrate more visual information and to make the decisions more robust. Hence, we extend our EIS with an elephant tracking approach such that detections of the same individual within consecutive images in a video are linked together to a so-called tracklet and all images of the individual within a tracklet can be used to perform the identification.

This time, we use the MegaDetector (Beery et al. 2019) for animal detection in the images, which has been pretrained on millions of camera trap images containing different animal species and which is built to detect animals in general without inferring the corresponding species. This general detector worked surprisingly well for both daytime and nighttime footage in our elephant dataset such that we did not need to carry out any additional training step with annotated elephant images.

[2] https://ccconservation.org/.

To connect detections to tracklets, we follow the tracking-by-detection approach and apply either the IOU/V-IOU tracker (Bochinski et al. 2017, 2018) or the graph-based multi-object tracking (GBMOT) approach (Mothes and Denzler 2017). For the identification, we then use every n-th image (for this experiment, we set $n = 15$) of each tracklet to reduce redundancy such that different images are distinct, and apply feature extraction and classification following our EIS from "The elephant identification system (EIS)". The choice of n is a trade-off between the amount of information or training data (lower n), and the distinctiveness of the individual examples (higher n). We also include the elephant head detector of our EIS, applied to the bounding boxes of each tracklet, and compare our EIS with the CurvRank approach of (Weideman et al. 2020) that uses features from the ear contours of the elephants for identification.

Results for identifying Asian elephants in videos

In preliminary experiments, we compared the different tracking approaches on ten annotated daytime videos and ten annotated nighttime videos of our elephant dataset. While the IOU/V-IOU tracker achieved higher precision (90.3% for daytime videos and 89.5% for nighttime videos) compared to GBMOT (85.8% for daytime videos and 88.8% for nighttime videos), the latter obtained a higher recall (79.5% for daytime and 83.2% for nighttime vs. 72.7% for daytime and 74.6% for nighttime). Hence, to not miss any individual, we decided to select the GBMOT approach for tracking due to its higher recall.

The results for identifying individual elephants are shown in Table 2. Across all videos, our EIS achieves a top-1 accuracy of 44.8% and a top-5 accuracy of 69.8% when considering the full bounding boxes of the elephants for identification. An additional elephant head detection did not change the results much, neither only for nighttime videos nor only for daytime videos. This can be attributed to the fact that the head detector has been trained on the other elephant dataset, which is of much higher quality compared to the images of the videos from the camera traps.

The low image quality is also the reason why CurvRank performed poorly in our experiments, because often no ear contour of the elephant could be found, or another contour not belonging to the ear was found that was then classified incorrectly. Note that the CurvRank algorithm was proposed for high quality photographs and in our experiments we have observed the limitations of this identification approach. However, as it is to the best of our knowledge the state-of-the art elephant-specific identification method, and our videos are of reasonably high resolution, CurvRank is nevertheless a sensible baseline.

This also highlights the difficulty of the dataset and emphasizes the good results of the EIS even more. Interestingly, the identification was consistently more successful on nighttime videos compared to daytime videos for both approaches.

In "Elephant identification results", we discuss the intended human-in-the-loop application of the EIS, which also applies to this video task. Consequently, the same limitation of a closed set applies here as well.

Identifying gorillas

In this section, we showcase an application (Brust et al. 2017) where our decoupled approach is a particularly good fit. The goal is to identify 147 individuals of the Western lowland gorilla species (*Gorilla gorilla gorilla*) using facial features.

Gorilla dataset

Instead of automated camera traps, the images are generated during manual field photography in the Nouabalé-Ndoki National Park, Republic of Congo. The photographers

Table 2 Our experimental results for identifying Asian elephants in video clips

	Full elephant bounding boxes		With detected elephant heads	
	Top-1 accuracy (%)	Top-5 accuracy (%)	Top-1 accuracy (%)	Top-5 accuracy (%)
All videos (175)				
CurvRank	4.5	17.2	4.9	12.4
EIS	44.8	69.8	44.3	69.6
Only nighttime videos (73)				
CurvRank	5.3	18.4	8.3	8.3
EIS	45.2	74.0	46.2	73.1
Only daytime videos (102)				
CurvRank	4.2	16.7	4.4	13.0
EIS	44.4	66.7	43.8	68.5

Fig. 4 Examples from the dataset of Brust et al. (2017) illustrating challenges such as occlusion, motion blur, lighting difficulties, and high variance in object scale

attempt to film individuals separately and each of the 12,765 images is labeled with the one individual in focus. However, there are no bounding box annotations for the respective faces. While the photographs are of high quality, there are numerous challenges as exemplified in Fig. 4.

Face detection

We randomly select 2500 images and manually annotate bounding boxes for each of the faces. The resulting small dataset is used to train a YOLO detector (Redmon et al. 2016) with a setup identical to the chimpanzee detector proposed by Freytag et al. (2016). We split the dataset into up to 2000 images for training and 500 for validation. Using all 2000 images for training, the detector achieves an average precision of 90.8%. With only 500 training images, we still reach a usable average precision of 86.6%.

As a sanity check, we run the detector on the whole dataset of 12,765 images and count the number of detections, assuming that each photograph contains exactly one individual. The detector trained on 2000 images detects exactly one face in 95.4% of images, no face in 0.4% and more than one face in 4.1%. Qualitative samples indicate that a large fraction of the false positives are in fact faces of infants that are sitting on an adult individual's shoulders.

Individual gorilla identification

We crop each face in the whole dataset using the detector and associate it with the respective labeled individual. This assumes that the face is detected correctly. When we detect more than one face, we select the face with the largest area because annotators tend to label the adult gorilla if an infant is present. We then extract features from the resulting dataset using the `pool5` layer of the BVLC AlexNet (Krizhevsky et al. 2012) implementation.[3]

The features are classified using an SVM similar to Freytag et al. (2016). Using the best detector trained on 2000 images, we achieve a top-1 identification accuracy of 62.4% and a top-5 accuracy of 80.3% over all 147 individuals.

The training process on the largest dataset completes in less than a second on modern x86 hardware. Hence, we do not have to rely on incremental learning in this task, but can afford to completely re-train the classifier whenever new training data becomes available. Including new individuals is trivial in this setup. For a proper evaluation of incremental learning techniques especially over a longer time period, large annotated image datasets from long-term monitoring

[3] http://dl.caffe.berkeleyvision.org/bvlc_alexnet.caffemodel.

studies are required, which are so far not available for many species including gorillas.

Identifying chimpanzees

In our final set of experiments, we focus on chimpanzees (*Pan troglodytes*) and start with the identification task in this section following Freytag et al. (2016). Afterwards, we consider attribute predictions in "Attribute predictions" and "Lifelong learning of age predictors". The dataset used in our experiments is described in the following.

C-Tai chimpanzee dataset

We use the C-Tai dataset described by Freytag et al. (2016), who also published the images as well as training and test splits for their experiments. The C-Tai dataset is derived from previously published datasets of Loos and Ernst (2013), who specifically focused on attribute predictions for chimpanzee faces. Hence, we also use it to evaluate our algorithms for the task of estimating attribute values of individuals.

The images of the C-Tai dataset have been recorded in the Tai National Park in Côte d'Ivoire, with strongly varying image qualities due to heavy illumination changes and individuals that are captured in large distances. There are 5078 chimpanzee faces from 78 individuals in this dataset in total, but only 4377 of them have complete annotations with respect to identity and further attributes (age, age group, gender), resulting in 62 different individuals from five age groups. For more details and statistics about this dataset, we refer to the work of Freytag et al. (2016) and the corresponding supplementary material.

Setup for identifying chimpanzees

To evaluate our identification approach, we use five random splits of the dataset following stratified sampling with 80% of the images for training and hold-out 20% for testing. The performance is measured using averaged class-wise recognition rates (ARR). Feature extraction is either done using the VGGFaces network of Parkhi et al. (2015) trained on the Labeled Faces in the Wild dataset (Huang et al. 2007) for human face recognition, or using the BVLC AlexNet (Krizhevsky et al. 2012) pre-trained on object categories of ImageNet (Russakovsky et al. 2015). For both networks, activation outputs of the `pool5` layer (last layer before fully-connected layers) and of the `fc7` layer (last layer before class scores) are tested. An SVM is used for classification (see "Classification and regression for individual identification and attribute prediction"). Furthermore, we compare our approach with the baseline of Loos and Ernst (2013) for chimpanzee identification.

Table 3 Our experimental results for identifying chimpanzees in the C-Tai dataset (Freytag et al. 2016)

Approach (features from)	ARR (%)
VGGFaces, `pool5`	68.0
VGGFaces, `fc7`	53.0
BVLC AlexNet, `pool5`	**76.6**
BVLC AlexNet, `fc7`	67.0
Baseline of Loos and Ernst (2013)	64.4

Best results indicated in bold

Results for identifying chimpanzees

Our experimental results for identifying chimpanzees in the C-Tai dataset are shown in Table 3. It can be observed that the quality of the identification heavily depends on the selected method for feature extraction. Hence, choosing an appropriate network architecture and a suitable activation layer is crucial for obtaining the best performance, which has been achieved by the `pool5` layer of the BVLC AlexNet with an average recognition rate of 76.6%. Note that three out of the four configurations shown in Table 3 outperform the baseline approach of Loos and Ernst (2013), which achieved 64.4% on the same experimental setup (Freytag et al. 2016).

Attribute predictions

Since the C-Tai dataset of chimpanzee faces described in "C-Tai chimpanzee dataset" contains attribute annotations, we have used this dataset for estimating attribute values of individuals using our general approach. In the following experiments, we consider the gender, the age group, and the age of the chimpanzees according to Freytag et al. (2016). We use the same experimental setup and feature representations that have been described in "Setup for identifying chimpanzees" for the identification task. Furthermore, we compare the results of our attribute prediction approach with a baseline method that performs the identification task as in the previous section and simply takes the attribute values of the predicted individual from the training set. Since the age sometimes changes for the same individual because images are recorded over multiple years, the baseline method takes the average age of the predicted individual calculated for the corresponding training set. This is done to obtain numerical results for the regression error when no meta data like the time stamps of the photos are used for the age prediction, thus only exploiting the pixel information of the images. Note that when using the time the photos were taken, one gets the age for free in case of correct identifications but when

Table 4 Experimental results for predicting attributes of individual chimpanzees in the C-Tai dataset (Freytag et al. 2016). The gender prediction is evaluated with the area under the ROC curve (AUC, higher is better), age group prediction is evaluated with average recognition rates (ARR, higher is better), and age prediction is evaluated with L_2-error (lower is better). For reference, we provide identification results as well

Approach (features from)	Gender prediction (AUC) (%)	Age group prediction (ARR) (%)	Age prediction (L_2-error)	Identification (ARR) (%)
VGGFaces, `pool5`	79.8	84.0	8.41	68.0
VGGFaces, `fc7`	88.0	76.4	8.35	53.0
BVLC AlexNet, `pool5`	**90.5**	**85.3**	6.79	**76.6**
BVLC AlexNet, `fc7`	87.0	83.8	**6.61**	67.0
Baseline (by identification)	89.6	77.9	8.30	76.6

Best result for each task indicated in bold

assigning the wrong ID, the age prediction error might even be larger compared to the averaging baseline. The results from our attribute prediction experiments are summarized in Table 4, and it can be observed that direct attribute prediction always leads to better results compared to the baseline of retrieving attributes via the identification. In the following, we go into the details for each considered attribute separately.

Gender prediction via binary classification

Gender prediction is a binary classification problem with two classes only, hence we train a linear binary SVM model for the different CNN features that we have extracted for the chimpanzee faces. We evaluate the performance of the binary classifiers using the area under the receiver operating characteristic (ROC) curve, in short area under the ROC curve (AUC), as a quantitative metric (Hanley and McNeil 1982; Fawcett 2006).

From the results in Table 4, we see that the baseline approach of retrieving the gender from the predicted individual already achieves a very good performance of 89.6% AUC. Hence, even if the wrong individual has been predicted for the identification task as indicated by the results in "Identifying chimpanzees", it has at least the same gender as the true individual. This suggests that if the individual classifier makes a mistake, it more likely confuses male chimpanzees with male chimpanzees and female chimpanzees with female chimpanzees. Nevertheless, slightly better performance for the gender prediction can be achieved by directly estimating the attribute value based on features from the `pool5` layer of the BVLC AlexNet, achieving the highest accuracy of 90.5% AUC.

Age group prediction via multi-class classification

As a second discrete attribute, we consider the age group of the chimpanzees as a rough estimate of age. For the C-Tai dataset, there are annotations for five age groups (classes) available: *Infant, Juvenile, SubAdult, Adult,* and *Elderly*. We refer to the original work of Freytag et al. (2016) for details about the distribution of individuals among these age groups. In our experiments, we split the examples of each age group into 90% training data and 10% hold-out test data, which is repeated five times to obtain reliable results. For classification into the five age groups, we train a linear multi-class SVM. The performance is measured by the average recognition rate (ARR) across the five age groups.

We compare the same feature extraction approaches as for gender prediction in the previous section and the results in Table 4 show that direct age group prediction leads to substantially better results with three out of the four used feature representations compared to the baseline approach of retrieving the attribute via identification. While the latter achieved an ARR of 77.9%, the best result has again been achieved by the BVLC AlexNet with features from the `pool5` layer (85.3%). For the age group prediction, features from the `pool5` layers were better suited compared to features from the `fc7` layers with both network architectures, indicating the importance of features from earlier layers within the network. When comparing architectures, the BVLC AlexNet trained for general object categories outperformed the VGGFaces trained for human face identification with respect to both feature representations from the chosen activation layers.

Age prediction via regression

The last attribute we consider for our general prediction approach is the age of the individuals. This is treated as a continuous variable for which we require a regression method to estimate suitable attribute values. We use GP regression (see "Classification and regression for individual identification and attribute prediction") for estimating the age, equipped with an RBF kernel as a covariance function and optimized hyperparameters as outlined by Freytag

et al. (2016). For training an age predictor, 100 images of individuals have been randomly selected, and the remaining images are used for hold-out testing. Random selection and model learning are repeated five times as in the previous experiments to obtain meaningful results. The performance of the age regression is measured by the L_2-error, which is the L_2-norm of the vector of residuals that contains the differences between the true values and the predicted values.

In the last column of Table 4, the averaged L_2-errors are shown for the different feature representations used in our general prediction approach. While features from the VGGFaces network perform similar to the baseline approach of retrieving attributes via identification (L_2 errors between 8.41 and 8.30), the features from the BVLC AlexNet are better suited for direct age prediction and using the `fc7` layer outputs results in the lowest L_2-error (6.61) achieved in our experiments. This highlights again the usefulness of features obtained from a deep neural network architecture that has been trained on images of common object categories (BVLC AlexNet pre-trained on ImageNet). Interestingly, in contrast to the two classification approaches used for the other two attributes in the previous sections, features of the `fc7` perform slightly better compared to the `pool5` layer outputs for the regression task.

Lifelong learning of age predictors

In our final experiment, we consider attribute prediction in a lifelong learning scenario and pick the regression task to look at continuous learning of age predictors (Käding et al. 2018). We use the EMOC criterion for active learning (see "Active learning strategies and incremental learning") to improve the model and its estimations over time by incorporating additional images automatically selected for annotation and model parameter update. Note that we simulate the data annotation process using the labels provided with the dataset in lieu of actual human annotations. The lifelong learning experiment is designed as follows.

Experimental setup for lifelong learning experiment

From 4414 images with age annotations of the C-Tai dataset ("C-Tai chimpanzee dataset"), only four are used for learning an initial GP regression model with RBF kernel as in "Age prediction via regression". The remaining 4410 images are split into 2205 instances for hold-out testing and 2205 instances that serve as the unlabeled pool for querying additional training data. We repeat the dataset splits three times and query 1000 samples sequentially in each experiment. After each query, the performance of the updated model is validated on the held-out test set and measured with the root-mean-square error (RMSE). As feature representations, we use L_2-normalized activation outputs from the `fc7` layer of the BVLC AlexNet, since these features performed best in our experiments for predicting the age of chimpanzees in "Age prediction via regression".

Active learning methods used for comparison

Besides our proposed EMOC strategy described in "Active learning strategies and incremental learning" for selecting query images in active learning, several competing methods are tested as well. A purely random selection of samples also known as passive learning (Yu and Kim 2010) is used as a baseline (`Random`). Furthermore, we consider selecting samples that either have the largest predictive variance (Kapoor et al. 2010) inferred from the GP regression model (`Variance`) or maximize the data entropy (`Entropy`). Exploration guided active learning (`EGAL`) of Hu et al. (2010) has been employed as well as several combinations of their introduced measurements for diversity (`Di`) and density (`Di`), denoted by `Di`λ/`De`$(1-\lambda)$ with $\lambda \in \{0.0, 0.5, 1.0\}$. Related to these measures is the baseline of querying samples with the largest Mahalanobis distance in feature space to already labeled samples (`Mahalanobis`). Finally, we compare EMOC with the expected model change (EMC) proposed by Cai et al. (2013).

Lifelong learning results

The results of the lifelong learning experiments for improving the age predictors on the C-Tai dataset are summarized in Fig. 5. We observe that the EMOC strategy reduces the prediction error (RMSE) the most, leading to the smallest area under the error curve. In fact, the GP regression model updated with samples selected by EMOC achieves the lowest errors after each number of added samples. Note that all competing methods are also outperformed by the `Random` baseline for passive learning, which indicates that appropriate sample selection for age prediction on the C-Tai dataset is a hard task. Nevertheless, EMOC leads to the highest model accuracies throughout the whole time span of the simulated lifelong learning scenario.

The average runtime of a single iteration including selection, model update, and predictions, is 497 ms for random selection (Käding 2020) on modern x86 hardware, averaged over the whole experiment as the runtime increases with the number of samples added to the underlying GP model. Using EMOC, it is 2040 ms. Still, the consistently strong reduction in annotation time when using EMOC should be considered as well, and strongly outweighs the small computational overhead. While other selection criteria, e.g. variance or entropy, have a negligible overhead, they do not perform as well as EMOC. Furthermore, EMOC can be approximated, e.g., following the strategy outlined in Käding et al. (2016a),

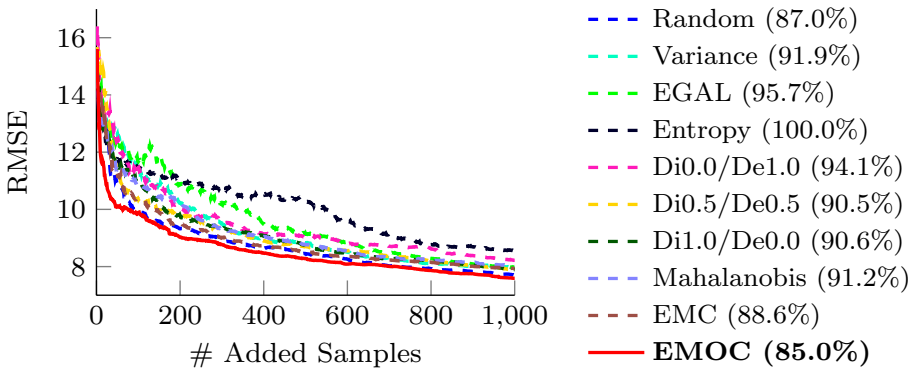

Fig. 5 Experimental results for lifelong learning with different active learning strategies. Error curves are shown, which indicate the RMSE evaluated after adding each of the 1000 samples to the training set and updating the prediction model. Numbers in the legend denote the area under the corresponding error curve (lower is better) relative to the worst performing method

or combined with random pre-selection (Brust et al. 2019) to reduce overhead.

We can conclude that our general approach for attribute prediction in combination with a suitable active learning strategy like EMOC is a good choice for lifelong learning, and clear model improvements can be achieved over time when adding additional data that become available during an application.

Conclusions

In this paper, we have shown that a general approach of using image features from pre-trained deep neural networks and decoupled decision models works well for identifying individuals in images and videos. This has been verified in experiments for identifying individuals of four different mammalian species, two elephant species (African and Asian) and two great apes (gorillas and chimpanzees). Although the achieved results are already remarkable when considering the varying challenges of the used datasets and the rather small amount of training data, there is still a lot of room for improvements to make the recognition systems even more valuable for practitioners.

We believe that the described concept of lifelong learning together with active learning and human-in-the-loop is able to achieve these improvements when it is applied during long-term monitoring studies with a continuous stream of new image data recorded over time. The targeted selection of the most relevant examples by active learning and the exploitation of expert knowledge through annotations via human-in-the-loop allows for continuous enhancements of the recognition models within the lifelong learning cycle, while at the same time reducing the human efforts for labeling additional data.

We have presented a way for integrating our general approach in a lifelong learning setup, highlighting the importance to exploit new incoming data during an application to improve the predictions of the models over time. In long-term monitoring applications, where additional data become available continuously, a conscious effort must be made to distribute human and computing resources evenly over time and maximize efficiency. Lifelong learning provides this distribution in two ways. First, active learning selects only important new data for annotation by human experts, whose time is often constrained. Second, incremental learning performs efficient and frequent model updates to cope with a potentially infinitely growing set of training data and provide a tight feedback loop together with active learning. Without lifelong learning, i.e., in a waterfall setting, a continuous data stream without intelligent selection by active learning will eventually overwhelm both the annotators and the available computing power. Hence, we believe that there is no alternative to a lifelong learning setup in long-term monitoring applications.

Besides its applicability for lifelong learning, the decoupling of image feature extraction from the final prediction task has further advantages. On the one hand, pre-trained deep neural networks can still be used to compute appropriate features for identification, and it is not necessary to learn network parameters solely based on data from the identification task. This is beneficial because initial labeled datasets for animal re-identification are often rather small at the beginning of a monitoring study, which makes optimizing large neural networks difficult. However, large networks are required for good performance because they allow for extracting semantically meaningful features, and exploiting pre-trained networks leverages the existing large-scale datasets of common object categories. Due to the decoupling of image feature extraction from the prediction task in our approach, these rich feature representations can be utilized and only the final decision model needs to be updated via efficient incremental learning algorithms within lifelong learning.

On the other hand, our approach allows for exchanging the final part of the processing pipeline, which is the decision model used to make the prediction. This can either be a classifier for assigning animal IDs, or a regression method for estimating continuous outputs such as the age of an individual. Thus, we are flexible in using the rich

feature representations from the deep neural networks for various prediction tasks and can select an appropriate decision model. By going beyond the standard task of identifying individuals, we have demonstrated the benefits of our approach to directly predict attributes of individuals. For the tasks of gender prediction, age group prediction, and age prediction, we have achieved superior results compared to the baseline approach of retrieving the corresponding attribute value from the identified individual after performing the identification. The binary classification problem of gender prediction as well as assigning one of five age groups has been tackled with support vector machine classifiers, whereas the age prediction has been performed with Gaussian process regression. Both classifiers and regression models can be continuously updated by efficient incremental learning techniques to further enhance the recognition system during its application via lifelong learning.

To summarize, we have provided a lifelong learning concept that is applicable for various monitoring tasks including individual identification and attribute prediction, and which exploits additional image data that becomes available over time. While pre-trained neural networks can also be leveraged for feature extraction in lifelong learning through our decoupled approach, the steady improvements of the decision models by incorporating expert feedback via active learning with human-in-the-loop lead to clear advantages compared to fixed recognition models that are trained only once on standard datasets before the application and are later kept unchanged. Hence, long-term monitoring of mammals based on image data can be further enhanced by implementing a lifelong learning cycle with a tight feedback loop that continuously incorporates expert knowledge during the whole application.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42991-022-00224-8.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Arandjelovic M, Head J, Kuehl H, Boesch C, Robbins MM, Maisels F, Vigilant L (2010) Effective non-invasive genetic monitoring of multiple wild western gorilla groups. Biol Cons 143(7):1780–1791. https://doi.org/10.1016/j.biocon.2010.04.030

Ardovini A, Cinque L, Sangineto E (2008) Identifying elephant photos by multi-curve matching. Pattern Recogn 41(6):1867–1877. https://doi.org/10.1016/j.patcog.2007.11.010

Bakliwal K, Ravela S (2020) The sloop system for individual animal identification with deep learning. arXiv preprint. arXiv:2003.00559

Beery S, Morris D, Yang S (2019) Efficient pipeline for camera trap image review. In: KDD Workshop on Data Mining and AI for Conservation. arXiv:1907.06772

Beery S, Van Horn G, Perona P (2018) Recognition in terra incognita. In: European Conference on Computer Vision (ECCV), pp 472–489. https://doi.org/10.1007/978-3-030-01270-0_28

Bendale A, Boult TE (2015) Towards open world recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1893–1902. https://doi.org/10.1109/CVPR.2015.7298799

Berger-Wolf TY, Rubenstein DI, Stewart CV, Holmberg JA, Parham J, Menon S, Crall J, Oast JV, Kiciman E, Joppa L (2017) Wildbook: crowdsourcing, computer vision, and data science for conservation. arXiv preprint. arXiv:1710.08880

Bochinski E, Eiselein V, Sikora T (2017) High-speed tracking-by-detection without using image information. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–6. https://doi.org/10.1109/AVSS.2017.8078516

Bochinski E, Senst T, Sikora T (2018) Extending iou based multi-object tracking by visual information. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–6. https://doi.org/10.1109/AVSS.2018.8639144

Bodesheim P, Freytag A, Rodner E, Denzler J (2015) Local novelty detection in multi-class recognition problems. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp 813–820. https://doi.org/10.1109/WACV.2015.113

Bodesheim P, Freytag A, Rodner E, Kemmler M, Denzler J (2013) Kernel null space methods for novelty detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3374–3381. https://doi.org/10.1109/CVPR.2013.433

Böhlke J, Korsch D, Bodesheim P, Denzler J (2021a) Exploiting web images for moth species classification. In: Computer Science for Biodiversity Workshop (CS4Biodiversity) at the Annual Conference of the German Informatics Society, (accepted for publication)

Böhlke J, Korsch D, Bodesheim P, Denzler J (2021b) Lightweight filtering of noisy web data: Augmenting fine-grained datasets with selected internet images. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp 466–477. https://doi.org/10.5220/0010244704660477

Branson S, Van Horn G, Belongie S, Perona P (2014) Improved bird species categorization using pose normalized deep convolutional nets. In: British Machine Vision Conference (BMVC). https://doi.org/10.5244/C.28.87

Brust CA, Barz B, Denzler J (2021) Carpe diem: A lifelong learning tool for automated wildlife surveillance. In: Computer Science for Biodiversity Workshop (CS4Biodiversity) at the Annual

Conference of the German Informatics Society, (accepted for publication)
Brust CA, Burghardt T, Groenenberg M, Käding C, Kühl H, Manguette ML, Denzler J (2017) Towards automated visual monitoring of individual gorillas in the wild. In: IEEE International Conference on Computer Vision Workshops (ICCVW), ICCV Workshop on Visual Wildlife Monitoring, pp 2820–2830. https://doi.org/10.1109/ICCVW.2017.333
Brust CA, Käding C, Denzler J (2019) Active learning for deep object detection. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp 181–190. https://doi.org/10.5220/0007248601810190
Brust CA, Käding C, Denzler J (2020) Active and incremental learning with weak supervision. Künstl Intell 34:165–180. https://doi.org/10.1007/s13218-020-00631-4
Cai W, Zhang Y, Zhou J (2013) Maximizing expected model change for active learning in regression. In: IEEE International Conference on Data Mining, pp 51–60. https://doi.org/10.1109/ICDM.2013.104
Carter SJ, Bell IP, Miller JJ, Gash PP (2014) Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. J Exp Mar Biol Ecol 452:105–110. https://doi.org/10.1016/j.jembe.2013.12.010
Castro FM, Marin-Jimenez MJ, Guil N, Schmid C, Alahari K (2018) End-to-end incremental learning. In: European Conference on Computer Vision (ECCV), pp 241–257. https://doi.org/10.1007/978-3-030-01258-8_15
Cauwenberghs G, Poggio T (2001) Incremental and decremental support vector machine learning. In: Conference on Advances in Neural Information Processing Systems (NIPS), pp 409–415
Cheema GS, Anand S (2017) Automatic detection and recognition of individuals in patterned species. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), pp 27–38. https://doi.org/10.1007/978-3-319-71273-4_3
Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1800–1807. https://doi.org/10.1109/CVPR.2017.195
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
Crall JP, Stewart CV, Berger-Wolf TY, Rubenstein DI, Sundaresan SR (2013) Hotspotter – patterned species instance recognition. In: IEEE Workshop on Applications of Computer Vision (WACV), pp 230–237. https://doi.org/10.1109/WACV.2013.6475023
Crunchant AS, Egerer M, Loos A, Burghardt T, Zuberbühler K, Corogenes K, Leinert V, Kulik L, Kühl HS (2017) Automated face detection for occurrence and occupancy estimation in chimpanzees. Am J Primatol 79(3):e22627. https://doi.org/10.1002/ajp.22627
Cui Y, Song Y, Sun C, Howard A, Belongie S (2018) Large scale fine-grained categorization and domain-specific transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4109–4118. https://doi.org/10.1109/cvpr.2018.00432
Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 886–893. https://doi.org/10.1109/CVPR.2005.177
Data Science Process Alliance (2021) What is waterfall? https://www.datascience-pm.com/waterfall/
Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
Diehl CP, Cauwenberghs G (2003) Svm incremental learning, adaptation and optimization. In: International Joint Conference on Neural Networks (IJCNN), pp 2685–2690. https://doi.org/10.1109/IJCNN.2003.1223991
Dunbar SG, Anger EC, Parham JR, Kingen C, Wright MK, Hayes CT, Safi S, Holmberg J, Salinas L, Baumbach DS (2021) Hotspotter: using a computer-driven photo-id application to identify sea turtles. J Exp Mar Biol Ecol 535:151490. https://doi.org/10.1016/j.jembe.2020.151490
Evans T, Brostow G, Jones K (2014) Active learning approaches to identifying animals in camera trap data. Tech. rep., CoMPLEX - UCL – University College London
Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010
Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell (TPAMI) 32(9):1627–1645. https://doi.org/10.1109/TPAMI.2009.167
French RM (1999) Catastrophic forgetting in connectionist networks. Trends Cogn Sci 3(4):128–135. https://doi.org/10.1016/S1364-6613(99)01294-2
Freytag A, Rodner E, Denzler J (2014) Selecting influential examples: active learning with expected model output changes. In: European Conference on Computer Vision (ECCV), pp 562–577. https://doi.org/10.1007/978-3-319-10593-2_37
Freytag A, Rodner E, Simon M, Loos A, Kühl HS, Denzler J (2016) Chimpanzee faces in the wild: log-euclidean CNNs for predicting identities and attributes of primates. In: German Conference on Pattern Recognition (GCPR), pp 51–63. https://doi.org/10.1007/978-3-319-45886-1_5
Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 317–326. https://doi.org/10.1109/CVPR.2016.41, arXiv:1511.06062
Ge W, Lin X, Yu Y (2019) Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3029–3038. https://doi.org/10.1109/CVPR.2019.00315
Girshick R (2015) Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV), pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 580–587. https://doi.org/10.1109/CVPR.2014.81
Guschanski K, Vigilant L, McNeilage A, Gray M, Kagoda E, Robbins MM (2009) Counting elusive animals: comparing field and genetic census of the entire mountain gorilla population of bwindi impenetrable national park, uganda. Biol Cons 142(2):290–300. https://doi.org/10.1016/j.biocon.2008.10.024
Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. https://doi.org/10.1148/radiology.143.1.7063747
Hayes TL, Kafle K, Shrestha R, Acharya M, Kanan C (2020) Remind your neural network to prevent catastrophic forgetting. In: European Conference on Computer Vision (ECCV), pp 466–483. https://doi.org/10.1007/978-3-030-58598-3_28
He X, Peng Y, Zhao J (2019) Which and how many regions to gaze: focus discriminative regions for fine-grained visual categorization. Int J Comput Vis (IJCV) 127:1235–1255. https://doi.org/10.1007/s11263-019-01176-2
He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV), pp 2980–2988. https://doi.org/10.1109/ICCV.2017.322

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90

Howe EJ, Buckland ST, Després-Einspenner ML, Kühl HS (2017) Distance sampling with camera traps. Methods Ecol Evol 8:1558–1565. https://doi.org/10.1111/2041-210X.12790

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Tech. Rep. 07-49. University of Massachusetts, Amherst

Hu R, Delany SJ, Namee BM (2010) EGAL: exploration guided active learning for TCBR. In: International Conference on Case-Based Reasoning (ICCBR), pp 156–170. https://doi.org/10.1007/978-3-642-14274-1_13

Käding C (2020) Human-in-the-loop: lifelong learning for shallow and deep models. PhD thesis, Friedrich-Schiller-Universität Jena

Käding C, Freytag A, Rodner E, Bodesheim P, Denzler J (2015) Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4343–4352. https://doi.org/10.1109/CVPR.2015.7299063

Käding C, Freytag A, Rodner E, Perino A, Denzler J (2016a) Large-scale active learning with approximated expected model output changes. In: German Conference on Pattern Recognition (GCPR), pp 179–191. https://doi.org/10.1007/978-3-319-45886-1_15

Käding C, Rodner E, Freytag A, Denzler J (2016b) Active and continuous exploration with deep neural networks and expected model output changes. In: Conference on Advances in Neural Information Processing Systems Workshops, NIPS Workshop on Continual Learning and Deep Networks. arXiv:1612.06129

Käding C, Rodner E, Freytag A, Denzler J (2016c) Fine-tuning deep neural networks in continuous learning scenarios. In: Asian Conference on Computer Vision Workshops, ACCV Workshop on Interpretation and Visualization of Deep Neural Nets, pp 588–605. https://doi.org/10.1007/978-3-319-54526-4_43

Käding C, Rodner E, Freytag A, Denzler J (2016d) Watch, ask, learn, and improve: a lifelong learning cycle for visual recognition. In: European Symposium on Artificial Neural Networks (ESANN), pp 381–386. https://www.esann.org/sites/default/files/proceedings/legacy/es2016-91.pdf

Käding C, Rodner E, Freytag A, Mothes O, Barz B, Denzler J, AG CZ (2018) Active learning for regression tasks with expected model output changes. In: British Machine Vision Conference (BMVC), p 103. http://www.bmva.org/bmvc/2018/contents/papers/0362.pdf

Kapoor A, Grauman K, Urtasun R, Darrell T (2010) Gaussian processes for object categorization. Int J Comput Vis (IJCV) 88(2):169–188. https://doi.org/10.1007/s11263-009-0268-3

Kellenberger B, Tuia D, Morris D (2020) Aide: accelerating image-based ecological surveys with interactive machine learning. Methods Ecol Evol 11(12):1716–1727. https://doi.org/10.1111/2041-210X.13489

Korsch D, Bodesheim P, Denzler J (2019) Classification-specific parts for improving fine-grained visual categorization. In: German Conference on Pattern Recognition (GCPR), pp 62–75. https://doi.org/10.1007/978-3-030-33676-9_5

Korsch D, Bodesheim P, Denzler J (2021a) Deep learning pipeline for automated visual moth monitoring: insect localization and species classification. In: Computer Science for Biodiversity Workshop (CS4Biodiversity) at the Annual Conference of the German Informatics Society, (accepted for publication)

Korsch D, Bodesheim P, Denzler J (2021b) End-to-end learning of fisher vector encodings for part features in fine-grained recognition. In: DAGM German Conference on Pattern Recognition (DAGM-GCPR), (accepted for publication)

Körschens M, Barz B, Denzler J (2018) Towards automatic identification of elephants in the wild. In: AI for Wildlife Conservation (AIWC) Workshop. arXiv:1812.04418

Körschens M, Denzler J (2019) Elpephants: a fine-grained dataset for elephant re-identification. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), ICCV Workshop on Computer Vision for Wildlife Conservation (CVWC), pp 263–270. https://doi.org/10.1109/ICCVW.2019.00035

Krause J, Sapp B, Howard A, Zhou H, Toshev A, Duerig T, Philbin J, Fei-Fei L (2016) The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision (ECCV), pp 301–320. https://doi.org/10.1007/978-3-319-46487-9_19

Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: IEEE Workshop on 3D Representation and Recognition (3dRR), pp 554–561. https://doi.org/10.1109/iccvw.2013.77

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Conference on Advances in Neural Information Processing Systems (NIPS), pp 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kühl H (2008) Best practice guidelines for the surveys and monitoring of great ape populations. 36, IUCN

Kühl HS, Burghardt T (2013) Animal biometrics: quantifying and detecting phenotypic appearance. Trends Ecol Evol 28(7):432–441. https://doi.org/10.1016/j.tree.2013.02.013

Kulits P, Wall J, Bedetti A, Henley M, Beery S (2021) Elephantbook: a semi-automated human-in-the-loop system for elephant re-identification. In: ACM SIGCAS Conference on Computing and Sustainable Societies, Association for Computing Machinery, New York, NY, USA, COMPASS '21, pp 88–98. https://doi.org/10.1145/3460112.3471947

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time-series. In: The Handbook of Brain Theory and Neural Networks

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1990) Handwritten digit recognition with a back-propagation network. In: Conference on Advances in Neural Information Processing Systems (NIPS), pp 396–404. https://papers.nips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html

Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp 3–12. https://doi.org/10.1007/978-1-4471-2099-5_1

Lin TY, RoyChowdhury A, Maji S (2015) Bilinear CNN models for fine-grained visual recognition. In: IEEE International Conference on Computer Vision (ICCV), pp 1449–1457. https://doi.org/10.1109/iccv.2015.170

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: European Conference on Computer Vision (ECCV), pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Liu C, Zhang R, Guo L (2019) Part-pose guided amur tiger re-identification. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), ICCV Workshop on Computer

Vision for Wildlife Conservation (CVWC), pp 315–322. https://doi.org/10.1109/ICCVW.2019.00042
Lomonaco V, Maltoni D (2017) Core50: a new dataset and benchmark for continuous object recognition. In: Annual Conference on Robot Learning, PMLR, Proceedings of Machine Learning Research, vol 78, pp 17–26. https://proceedings.mlr.press/v78/lomonaco17a.html
Long M, Cao Y, Cao Z, Wang J, Jordan MI (2019) Transferable representation learning with deep adaptation networks. IEEE Trans Pattern Anal Mach Intell (TPAMI) 41(12):3071–3085. https://doi.org/10.1109/TPAMI.2018.2868685
Loos A (2012) Identification of great apes using gabor features and locality preserving projections. In: Workshop on Multimedia Analysis for Ecological Data, pp 19–24. https://doi.org/10.1145/2390832.2390838
Loos A, Ernst A (2013) An automated chimpanzee identification system using face detection and recognition. EURASIP J Image Video Process 1:49. https://doi.org/10.1186/1687-5281-2013-49
Loos A, Pfitzer M, Aporius L (2011) Identification of great apes using face recognition. In: European Signal Processing Conference, pp 922–926. https://ieeexplore.ieee.org/document/7074032
Lütz A, Rodner E, Denzler J (2013) I want to know more: efficient multi-class incremental learning using gaussian processes. Pattern Recogn Image Anal Adv Math Theory Appl (PRIA) 23:402–407. https://doi.org/10.1134/S1054661813030103
Maltoni D, Lomonaco V (2019) Continuous learning in single-incremental-task scenarios. Neural Netw 116:56–73. https://doi.org/10.1016/j.neunet.2019.03.010
Matan O, Kiang RK, Stenard CE, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD, Lecun Y (1990) Handwritten character recognition using neural network architectures. In: 4th USPS Advanced Technology Conference, pp 1003–1011. http://yann.lecun.com/exdb/publis/pdf/matan-90.pdf
McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: the sequential learning problem. Psychology of Learning and Motivation, vol 24, Academic Press, pp 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8
Moskvyak O, Maire F, Dayoub F, Baktashmotlagh M (2020) Learning landmark guided embeddings for animal re-identification. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), WACV Workshop on Deep Learning Methods and Applications for Animal Re-Identification, pp 12–19. https://doi.org/10.1109/WACVW50321.2020.9096932
Mothes O, Denzler J (2017) Anatomical landmark tracking by one-shot learned priors for augmented active appearance models. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp 246–254. https://doi.org/10.5220/0006133302460254
Nepovinnykh E, Eerola T, Kalviainen H (2020) Siamese network based pelage pattern matching for ringed seal re-identification. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), WACV Workshop on Deep Learning Methods and Applications for Animal Re-Identification, pp 25–34. https://doi.org/10.1109/WACVW50321.2020.9096935
Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc Natl Acad Sci 115(25):E5716–E5725. https://doi.org/10.1073/pnas.1719367115
Norouzzadeh MS, Morris D, Beery S, Joshi N, Jojic N, Clune J (2020) A deep active learning system for species identification and counting in camera trap images. Methods Ecol Evol 12(1):150–161. https://doi.org/10.1111/2041-210X.13504
Parham J, Stewart C, Crall J, Rubenstein D, Holmberg J, Berger-Wolf T (2018) An animal detection pipeline for identification. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp 1075–1083. https://doi.org/10.1109/WACV.2018.00123
Parkhi OM, Vedaldi A, Zisserman A, et al. (2015) Deep face recognition. In: British Machine Vision Conference (BMVC), pp 41.1–41.12. https://doi.org/10.5244/C.29.41
Pebsworth PA, LaFleur M (2014) Advancing primate research and conservation through the use of camera traps: introduction to the special issue. Int J Primatol 35(5):825–840. https://doi.org/10.1007/s10764-014-9802-4
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, http://www.gaussianprocess.org/gpml/
Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH (2017) iCaRL: incremental classifier and representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5533–5542. https://doi.org/10.1109/CVPR.2017.587
Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 779–788. https://doi.org/10.1109/CVPR.2016.91
Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6517–6525. https://doi.org/10.1109/CVPR.2017.690
Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Conference on Advances in Neural Information Processing Systems (NIPS). https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html
Robins A (1993) Catastrophic forgetting in neural networks: the role of rehearsal mechanisms. In: New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, pp 65–68. https://doi.org/10.1109/ANNES.1993.323080
Rodner E, Freytag A, Bodesheim P, Fröhlich B, Denzler J (2017) Large-scale gaussian process inference with generalized histogram intersection kernels for visual recognition tasks. Int J Comput Vis (IJCV) 121:253–280. https://doi.org/10.1007/s11263-016-0929-y
Rodner E, Simon M, Brehm G, Pietsch S, Wägele JW, Denzler J (2015) Fine-grained recognition datasets for biodiversity analysis. In: CVPR Workshop on Fine-grained Visual Classification (FGVC). arXiv:1507.00913
Rodner E, Simon M, Fisher B, Denzler J (2016) Fine-grained recognition in the noisy wild: sensitivity analysis of convolutional neural networks approaches. In: British Machine Vision Conference (BMVC), pp 60.1–60.13. https://doi.org/10.5244/C.30.60
Roy J, Vigilant L, Gray M, Wright E, Kato R, Kabano P, Basabose A, Tibenda E, Kühl HS, Robbins MM (2014) Challenges in the use of genetic mark-recapture to estimate the population size of bwindi mountain gorillas (gorilla beringei beringei). Biol Cons 180:249–261. https://doi.org/10.1016/j.biocon.2014.10.011
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV) 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
Scheirer WJ, Rocha A, Sapkota A, Boult TE (2013) Towards open set recognition. IEEE Trans Pattern Anal Mach Intell (TPAMI) 35(7):1757–1772. https://doi.org/10.1109/TPAMI.2012.256
Scheirer WJ, Jain LP, Boult TE (2014) Probability models for open set recognition. IEEE Trans Pattern Anal Mach Intell (TPAMI) 36(11):2317–2324. https://doi.org/10.1109/TPAMI.2014.2321392

Schneider S, Greenberg S, Taylor GW, Kremer SC (2020) Three critical factors affecting automated image species recognition performance for camera traps. Ecol Evol 10(7):3503–3517. https://doi.org/10.1002/ece3.6147

Schneider S, Taylor GW, Kremer SC (2020b) Similarity learning networks for animal individual re-identification - beyond the capabilities of a human observer. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), WACV Workshop on Deep Learning Methods and Applications for Animal Re-Identification, pp 44–52. https://doi.org/10.1109/WACVW50321.2020.9096925

Schneider S, Taylor GW, Linquist S, Kremer SC (2019) Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods Ecol Evolut 10(4):461–470. https://doi.org/10.1111/2041-210X.13133, https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13133

Schölkopf B, Smola AJ (2001) Learning with Kernels: support vector machines, regularization, optimization, and beyond. The MIT Press. https://mitpress.mit.edu/books/learning-kernels

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Settles B (2009) Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences. http://burrsettles.com/pub/settles.activelearning.pdf

Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), CVPR Workshop on Deep learning in Computer Vision (DeepVision). https://openaccess.thecvf.com/content_cvpr_workshops_2014/W15/html/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.html

Shmelkov K, Schmid C, Alahari K (2017) Incremental learning of object detectors without catastrophic forgetting. In: IEEE International Conference on Computer Vision (ICCV), pp 3420–3429. https://doi.org/10.1109/ICCV.2017.368

Shukla A, anderson c, Sigh Cheema G, Gao P, Onda S, Anshumaan D, Anand S, Farrell R (2019) A hybrid approach to tiger re-identification. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), ICCV Workshop on Computer Vision for Wildlife Conservation (CVWC), pp 294–301. https://doi.org/10.1109/ICCVW.2019.00039

Simon M, Rodner E, Darell T, Denzler J (2020) The whole is more than its parts? From explicit to implicit pose normalization. IEEE Trans Pattern Anal Mach Intell (TPAMI) 42(3):749–763. https://doi.org/10.1109/TPAMI.2018.2885764

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). arXiv:1409.1556

Stewart CV, Parham JR, Holmberg J, Berger-Wolf TY (2021) The animal id problem: continual curation. arXiv preprint. arXiv:2106.10377

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tabak MA, Norouzzadeh MS, Wolfson DW, Sweeney SJ, Vercauteren KC, Snow NP, Halseth JM, Di Salvo PA, Lewis JS, White MD, Teton B, Beasley JC, Schlichting PE, Boughton RK, Wight B, Newkirk ES, Ivan JS, Odell EA, Brook RK, Lukacs PM, Moeller AK, Mandeville EG, Clune J, Miller RS (2019) Machine learning to classify animal species in camera trap images: applications in ecology. Methods Ecol Evol 10(4):585–590. https://doi.org/10.1111/2041-210X.13120

Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Villa AG, Salazar A, Vargas F (2017) Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Eco Inform 41:24–32. https://doi.org/10.1016/j.ecoinf.2017.07.004

Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-UCSD birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology. http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

Wang K, Zhang D, Li Y, Zhang R, Lin L (2017) Cost-effective active learning for deep image classification. IEEE Trans Circuits Syst Video Technol 27(12):2591–2600. https://doi.org/10.1109/TCSVT.2016.2589879

Weideman H, Stewart C, Parham J, Holmberg J, Flynn K, Calambokidis J, Paul DB, Bedetti A, Henley M, Pope F, Lepirei J (2020) Extracting identifying contours for African elephants and humpback whales using a learned appearance model. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 1265–1274. https://doi.org/10.1109/WACV45572.2020.9093266

Willi M, Pitman RT, Cardoso AW, Locke C, Swanson A, Boyer A, Veldthuis M, Fortson L (2019) Identifying animal species in camera trap images using deep learning and citizen science. Methods Ecol Evol 10:80–91. https://doi.org/10.1111/2041-210X.13099

Xie S, Girshick R, Dollar P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5987–5995. https://doi.org/10.1109/CVPR.2017.634

Yang X, Mirmehdi M, Burghardt T (2019) Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshop on Computer Vision for Wildlife Conservation (CVWC), pp 255–262. https://doi.org/10.1109/ICCVW.2019.00034

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Conference on Advances in Neural Information Processing Systems (NIPS), pp 3320–3328. https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html

Yu H, Kim S (2010) Passive sampling for regression. In: IEEE International Conference on Data Mining, pp 1151–1156. https://doi.org/10.1109/ICDM.2010.9

Yu J, Su H, Liu J, Yang Z, Zhang Z, Zhu Y, Yang L, Jiao B (2019) A strong baseline for tiger re-id and its bag of tricks. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), ICCV Workshop on Computer Vision for Wildlife Conservation (CVWC), pp 302–309. https://doi.org/10.1109/ICCVW.2019.00040

Zhang L, Huang S, Liu W, Tao D (2019) Learning a mixture of granularity-specific experts for fine-grained categorization. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp 8330–8339. https://doi.org/10.1109/ICCV.2019.00842