

ORIGINAL RESEARCH

Species-level image classification with convolutional neural network enables insect identification from habitus images

Oskar L. P. Hansen^{1,2,3}  | Jens-Christian Svenning^{1,2}  | Kent Olsen³  |
Steen Dupont⁴  | Beulah H. Garner⁴  | Alexandros Iosifidis⁵  |
Benjamin W. Price⁴  | Toke T. Høye⁶ 

¹Department of Bioscience, Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Aarhus University, Aarhus C, Denmark

²Department of Bioscience, Section for Ecoinformatics and Biodiversity, Aarhus University, Aarhus C, Denmark

³Natural History Museum Aarhus, Aarhus C, Denmark

⁴Life Sciences, Natural History Museum, London, UK

⁵Department of Engineering - Signal Processing, Aarhus University, Aarhus N, Denmark

⁶Department of Bioscience and Arctic Research Centre, Aarhus University, Rønde, Denmark

Correspondence

Oskar L. P. Hansen, Department of Bioscience, Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark.
Email: oli@bios.au.dk

Funding information

15. Juni Fonden, Grant/Award Number: 2017-N-10; Danish Agency for Culture and Palaces under the Danish Ministry of Culture, Grant/Award Number: FORM.2016-0025; Villum Fonden, Grant/Award Number: 16549 and 17523; Innovation Fund Denmark, Grant/Award Number: 6171-00034B; Carlsberg Foundation Semper Ardens project MegaPast2Future, Grant/Award Number: CF16-0005

Abstract

1. Changes in insect biomass, abundance, and diversity are challenging to track at sufficient spatial, temporal, and taxonomic resolution. Camera traps can capture habitus images of ground-dwelling insects. However, currently sampling involves manually detecting and identifying specimens. Here, we test whether a convolutional neural network (CNN) can classify habitus images of ground beetles to species level, and estimate how correct classification relates to body size, number of species inside genera, and species identity.
2. We created an image database of 65,841 museum specimens comprising 361 carabid beetle species from the British Isles and fine-tuned the parameters of a pretrained CNN from a training dataset. By summing up class confidence values within genus, tribe, and subfamily and setting a confidence threshold, we trade-off between classification accuracy, precision, and recall and taxonomic resolution.
3. The CNN classified 51.9% of 19,164 test images correctly to species level and 74.9% to genus level. Average classification recall on species level was 50.7%. Applying a threshold of 0.5 increased the average classification recall to 74.6% at the expense of taxonomic resolution. Higher top value from the output layer and larger sized species were more often classified correctly, as were images of species in genera with few species.
4. Fine-tuning enabled us to classify images with a high mean recall for the whole test dataset to species or higher taxonomic levels, however, with high variability. This indicates that some species are more difficult to identify because of properties such as their body size or the number of related species.
5. Together, species-level image classification of arthropods from museum collections and ecological monitoring can substantially increase the amount of occurrence data that can feasibly be collected. These tools thus provide new opportunities in understanding and predicting ecological responses to environmental change.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

KEYWORDS

arthropod sampling, automatic species identification, camera trap, entomological collection, image classification, image database

1 | INTRODUCTION

Recent reports suggest that insect biomass and abundance have been declining dramatically in recent decades (Agrawal & Inamine, 2018; Hallmann et al., 2017; Lister & Garcia, 2018; Loboda, Savage, Buddle, Schmidt, & Høye, 2018; Seibold et al., 2019; Wagner, 2019), even though trends vary if measured across or on individual habitats and species (Loboda et al., 2018). Estimating and tracking changes in abundance and diversity of insects at species level through time and space is critical to understand the underlying drivers of change and to devise possible mitigation strategies. Methods that enable error estimation in observations, with high data quantity, quality, and resolution on spatial, temporal and taxonomic scales are crucial.

To date, no efficient method enables tracking of insect activity, abundance, and diversity in a nondestructive, cost-effective, and standardized way. Common sampling methods including direct observations, a variety of trapping methods, direct sampling methods, and DNA-based methods all fail on one or two of these criteria. A much criticized but widely used method is pitfall traps (Brown & Matthews, 2016; Engel et al., 2017; Skvarla, Larson, & Dowling, 2014). Like other trapping methods such as malaise traps and pan traps, they remove study specimens from the environment, thus being invasive. Furthermore, each trapping method comes with its own set of biases or methodological idiosyncratic behaviors, making interpretations across habitats difficult (Skvarla et al., 2014). Given the sampling method and in order to increase the number of individuals trapped this often comes at the expense of coarse temporal information (several days or weeks; Schirmel, Lenze, Katzmann, & Buchholz, 2010). The resulting low temporal resolution in activity estimate defined by the sampling frequency can only be related to environmental factors over the same time scale (Asmus et al., 2018; Høye & Forchhammer, 2008). Direct observations, being nondestructive, currently require identification of organisms by trained ecologists or taxonomists at the study site throughout the sampling period, greatly reducing the number of feasible samples.

The camera trap method has distinct advantages over traditional methods in entomology. Compared to the often used pitfall traps, camera traps sample more individuals (Collett & Fisher, 2017; Halsall & Wratten, 1988), and cause no depletion of specimens or habitat destruction (Digweed, Currie, Carcamo, & Spence, 1995; Zaller et al., 2015). Furthermore, camera traps require less maintenance (Caravaggi et al., 2017; Collett & Fisher, 2017). The average movement speed and various behavioral traits of a species can be directly measured between single frames of one camera trap (Caravaggi et al., 2017), allowing true abundance of species to be estimated based on their movement speed and range. Rarely, but increasingly, camera traps have been used to monitor insects and other arthropods (Collett & Fisher, 2017; Dolek & Georgi, 2017; Zaller et al., 2015).

Even though identifications of species based on images are well known for mammals and birds (Norouzzadeh et al., 2018; Yu et al., 2013), camera trap studies designed for arthropods conclude that image-based species identification by humans is generally not possible (Collett & Fisher, 2017; Zaller et al., 2015).

Image-based species identification methods on arthropods have been applied with success on samples in the laboratory (Joutsijoki et al., 2014). In order to fully implement the advantages of camera traps, there is a need for implementing image classification techniques to automatically identify and recognize species (Weinstein, 2017). Deep convolutional neural networks have together with the release of machine learning frameworks like TensorFlow (Abadi et al., 2015) and available models like Inception or GoogleNet (Szegedy et al., 2015; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) have advanced significantly in recent years (Wäldchen & Mäder, 2018). Image classifications used for species identification have dramatically increased in accuracy, performance, and in the number of taxa analyzed (Marques et al., 2018; Martineau et al., 2017; Norouzzadeh et al., 2018; Schneider, Taylor, & Kremer, 2018; Van Horn et al., 2017). On a limited number of species, identification by computers can be as good as human experts and with less variation in accuracy (Årje et al., 2019). Automated species identification has also been successfully implemented on the citizen science portal iNaturalist.org, enabling a suggested list of species for an observation, based on the existing archive of image data (Van Horn et al., 2017).

We test the ability of a convolutional neural network (CNN) to classify ground beetles (Coleoptera: Carabidae) to genus, species, or higher taxonomic level from images of specimens within the British collection at the Natural History Museum, London. This collection provides a good test case as it has been well curated and assessed for correct species identity, represents a commonly prepared type of insect collection for which this method is directly applicable to, and has access to the SatScan[®] (SmartDrive Limited; Blagoderov, Kitching, Livermore, Simonsen, & Smith, 2012; Mantle, LaSalle, & Fisher, 2012), a rapid whole drawer imaging system. Beetle specimens are placed in unit trays inside drawers, prepared either glued onto card or pinned and are generally positioned in dorsal view with head in the same direction, reducing the variability in the data. These prepared specimens can serve a simplified model for what a camera trap would record. Thus, these images represent a good indicator of the potential taxonomic resolution of automatic species identification with current state of the art classification methods, based on data from a camera trap, when compared to expert identifications of the specimens. Specifically, we quantify the number of correct species identifications of carabid beetles based on image classification of habitus images. Furthermore, we assess variation in correctly classified images among taxa. In particular, we test how classification recall (number of images classified to a group from the total number of images within the group) varies among genera

and for specimens of different body size. To increase accuracy and to critically assess reliability, we postprocess the output and apply thresholds on confidence values for each of the included taxonomic levels to avoid low confidence in predictions.

2 | MATERIALS AND METHODS

2.1 | Obtaining images

In August 2017, we scanned the British collection of ground beetles (Coleoptera: Carabidae) at the Natural History Museum London using the SatScan® (Blagoderov et al., 2012; Mantle et al., 2012). The collection comprised 207 drawers with specimens curated and identified to species level (Figure 1). All drawers were scanned with the same light and exposure settings following the imaging protocol

described by Blagoderov et al. (2012) and resulted in images of $15,828 \times 15,565$ pixels (36 pixels/mm) per drawer.

Drawer images were segmented into specimens using Inselect version 0.1.36 (Hudson et al., 2015) followed by manual refinement by two people, resulting in 65,841 single-specimen images (per species mean = 182, range = 1–892). To reduce variability and avoid images of exceptional preparations, specimens mounted with dorsal side facing down, or with head, pronotum, or elytra missing, and larvae were tagged during the manual quality check and refinement step. Each specimen image was also tagged with the taxon name (genus and species), according to the collection data (361 taxa). We excluded specimens without a taxon name (66 specimens) or without proper identification to species level (100 specimens), larvae (27 specimens) and specimens mounted with dorsal side downward (296 specimens) or missing either the head, pronotum, or elytra (504 specimens). In order to secure sufficient image data to test the classification success,

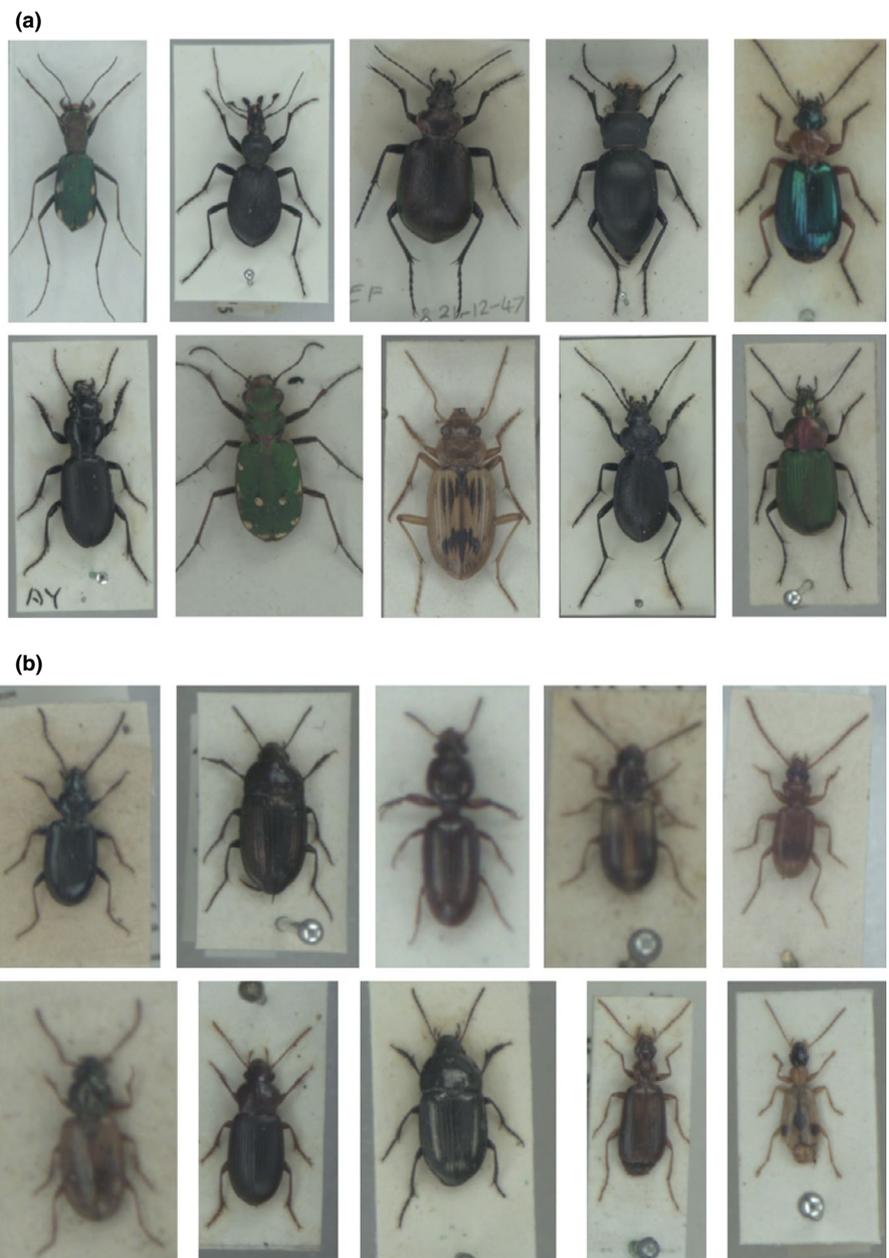


FIGURE 1 Specimens used to train or test the convolutional neural network. (a) Species with accuracy (images of species in test dataset classified to correct species) on 90% or more. First row: *Cylindera germanica*, *Cychnus caraboides*, *Calosoma inquisitor*, *Carabus glabratus*, and *Lebia chlorocephala*. Second row: *Brosicus cephalotes*, *Cicindela campestris*, *Nebria complanata*, *Carabus problematicus*, and *Chlaenius nigricornis*. (b) Species drawn randomly from the remaining 281 species. First row: *Bembidion atrocaeruleum*, *Amara famelica*, *Dyschirius politus*, *Acupalpus meridianus*, and *Blemus discus*. Second row: *Bembidion ephippium*, *Ophonus melletii*, *Amara lunicollis*, *Dromius angustus*, and *Demetrias imperialis*

only species with 50 specimens or more were included, thus excluding additional 70 species and 1,550 specimens. The taxonomic classification used for the species was from gbif.org via the *taxize* R-package (Chamberlain & Szöcs, 2013). Afterward additional taxonomic levels were added such as family, subfamily, tribe, subgenus, and the ordered taxonomic hierarchy from the British checklist of beetles (Duff, Lott, Buckland, & Buckland, 2012).

2.2 | Training and testing the convolutional neural network

The complete dataset comprised 63,364 specimen images from 291 species (images per species mean = 218, range = 50–888; Figure 1) comprising 80 genera. For each species, specimen images were divided into three groups for training (50%), validating (20%), and testing (30%) the network, respectively. In order to assign images consistently to the three datasets, we generated a probability value for each image based on the output from encrypting the filename. Images with percentage 0–20 were assigned as validation, 20–50 as testing and above 50 as training. Thus, the division percentages did not entirely reflect the number of images in each of the datasets with 31,533 (49.8%), 25,334 (20.0%), and 19,164 (30.2%) images used for training, validation, and testing, respectively. While the training and validation images were used only for training of the model, the test images, not known by the retrained model, were used for further analysis. We used the scripts developed by TensorFlow (Abadi et al., 2015) for training an Inception-v3 model (Szegedy et al., 2016) initially trained on ImageNet database (Deng et al., 2009), following the tutorial from Tensorflow (Tensorflow, 2019). The retraining was run in TensorFlow version 1.13.1, python version 3.7.3. Input images were resized to 299 × 299 pixels regardless of input image size and shape to follow the model specification. The model was trained with gradient decent optimizer for 225,000 iterations and a batch size of 100 for both training and validation datasets to reach at least 700 epochs. We did not apply augmentation of images; however, we tested that the learning rate of the model was optimized by training the model with learning rates of 0.5, 0.3, 0.1,

0.045, 0.01, 0.001, and 0.0001. We choose the default learning rate, that is, 0.045, which produced similar validation accuracy as 0.5, 0.3, and 0.1 with lower learning rate. Thus choosing the hyperparameter with smallest optimization update while the validation accuracy converged during training steps. The output layer of the CNN, activated by a softmax function, gave a predicted confidence value for each of the 291 species in each image ranging between 0 and 1.

2.3 | Evaluating predictions and setting thresholds to separate low- and high-confidence predictions

The output layer of the convolutional neural network consisted of a vector with a confidence value for each class (i.e., species) included in the neural network. The class with the highest value (top 1 or top 5) from the output layer was interpreted as the predicted class for an image. We assessed if the image was predicted correctly, if the ground truth name appeared in top 1 or, in a separate measure, in top 5.

For each species we calculated from all test images, true positives (tp): ground truth and predicted as ground truth, false positives (fp): not ground truth and predicted as ground truth, true negative (tn): not ground truth and not predicted as ground truth, false negative (fn): ground truth and not predicted as ground truth. Based on these numbers, we evaluated classification precision: $tp/(tp + fp)$, classification recall: $tp/(tp + tn)$, classification accuracy: $(tp + tn)/(tp + tn + fp + fn)$, True positive rate (TPR): $tp/(tp + fn)$, true negative rate (TNR): $tn/(tn + fp)$, and balanced classification accuracy: $(TPR + TNR)/2$.

The neural network included only species-level classes. To assess the number of correctly classified images on levels above the species level, we calculated a new set of confidence values through the sum of all classes in the higher taxonomic level (e.g., the confidence value sum of all species belonging to the same genus). We repeated this procedure for all taxonomic levels (subgenus, genus, tribe, subfamily, and family).

We introduced a minimum confidence value threshold to assess at which taxonomic resolution an image could be classified. Starting at

Taxon rank	No. specimens	Mean balanced accuracy	Mean precision	Mean recall
Species	10,348	82.4	70.0	64.9
Subgenus	1,799	73.1	75.3	46.5
Genus	3,212	65.0	72.3	30.6
Tribe	1,231	64.1	72.0	30.1
Subfamily	2,573	70.0	66.6	66.6
Family	1	NA	100	100
Weighted mean	19,164	75.8	70.6	55.4

Note: Bottom row gives the mean of measures weighted by the number of specimens at each taxon rank.

TABLE 1 Number of specimens at each taxonomic resolution, mean balanced accuracy, mean precision, and mean recall when setting a minimum acceptable confidence threshold to 0.5 before decreasing taxonomic resolution

species-level resolution, we evaluated if the highest confidence value was below the threshold value. If the highest confidence value was lower than the threshold value, we repeated the evaluation for classes at the next taxonomic level, that is, at lower taxonomic resolution.

2.4 | Analysis

In total 19,164 images of 291 species (mean images per species = 65.9, range = 11–272) were used as test images, not involved in the training and validation. As the number of images was not equal for all species, classification recall was calculated for each species as the proportion of images correctly classified. We used two generalized linear models with binomial distribution to assess if a classification of an image was correct or not. In model 1, only species identity was used as explanatory variable. In model 2, we used image size measured in megapixels extracted from the image metadata (exif) using *exiftool* v.11.06 through *exifr* r-package (Dunnington & Harvey, 2019) as a measure of body size (hereafter referred to as body size), the number of species within its genus, and the top 1 value from the output layer in the convolutional neural network as explanatory variables. A sensitivity analysis was performed for model 2, to separate effects from the three explanatory variables on the prediction. This analysis kept all but one variable constant at mean value for number of training images, body size, and top1 value or median for number of species inside genus.

3 | RESULTS

Of the 19,164 test images, 9,949 (51.9%) were predicted to the correct species (threshold = 0, n species = 291), while when extracting genus names of predictions and ground truths 14,357 (74.9%) were

predicted to the correct genus (n genera = 80). For predictions on species level, mean classification precision, recall, accuracy, and balanced accuracy were 54.7%, 50.7%, 99.7%, and 75.3%, respectively (Table S1).

The confusion matrix, based on all species-level predictions, revealed that species were often confused with other species within the same genus (Figure S1). Some typical confusions were also between tribes: Bembidiini species were often mistaken as Lebiini species and vice versa. On the other hand, Bembidiini were rarely mistaken as Zabryni, while Zabryni were mistaken as Bembidiini in more images. On the subfamily level, two tiger beetles (Cicindelinae) were misidentified as belonging to one of the other subfamilies, while six species of Carabinae had at least one image predicted as Cicindelinae (Figure S1).

By excluding images with low confidence (<0.25) on any one specific species (top 1 value), 56.6% and 77.8% of a total of 16,812 images were predicted to the correct species and genus, respectively. The resulting mean classification precision, recall, accuracy, and balanced accuracy were 58.2%, 54.0%, 99.7%, and 76.9%, respectively. Average recall per species was 50.7% (min 10.6, max 100%, SD 20.8%, SE 1.2%; Figure S2).

Setting a minimum acceptable confidence threshold to 0.5 before decreasing taxonomic resolution by one hierarchical level (i.e., summing all species-level confidence values from species belonging to that group e.g., all species in a genus), 75.8% of a total of 19,164 images were classified correctly to the decided taxonomic level and average classification recall across all specimens increased to 74.6% (min 21.3%, max 98.2%, SD 13.2%, SE 0.8%). Mean balanced accuracy, precision, and recall varied with taxonomic resolution (Table 1). Classification recall and taxonomic resolution varied considerably among the 291 species and 80 genera (Figure 2). For most species, the proportion of images correctly classified was above 76.9% (median, Figure 2a). Of the 14,527

TABLE 2 Species with 90% or more of specimens predicted to correct species if no threshold was set

Species	Number of specimens	No threshold		Threshold 0.5		
		True positive (%)	False positive	True positive (%)	Not species level	False positive
<i>Cylindera germanica</i>	39	100	12	94.9	2	4
<i>Cychrus caraboides</i>	57	98.2	13	96.5	2	4
<i>Calosoma inquisitor</i>	39	97.4	8	97.4	0	3
<i>Carabus glabratus</i>	27	96.3	2	88.9	2	NA
<i>Lebia chlorocephala</i>	67	95.5	3	94.0	1	3
<i>Broscus cephalotes</i>	111	93.7	14	92.8	3	4
<i>Cicindela campestris</i>	88	93.2	NA	92.0	3	NA
<i>Nebria complanata</i>	55	92.7	1	89.1	4	NA
<i>Carabus problematicus</i>	107	91.6	16	87.9	7	14
<i>Chlaenius nigricornis</i>	40	90.0	4	82.5	5	3

Note: Number of specimens in test dataset, percentage of specimens predicted to the correct species, and false positives (i.e., the number of specimens predicted to the wrong species). For threshold 0.5, the percentage of specimens in species predicted to the correct species, the number of specimens that did not meet the threshold, thus not predicted on species level, and false positives (i.e., the number of specimens predicted to the wrong species).

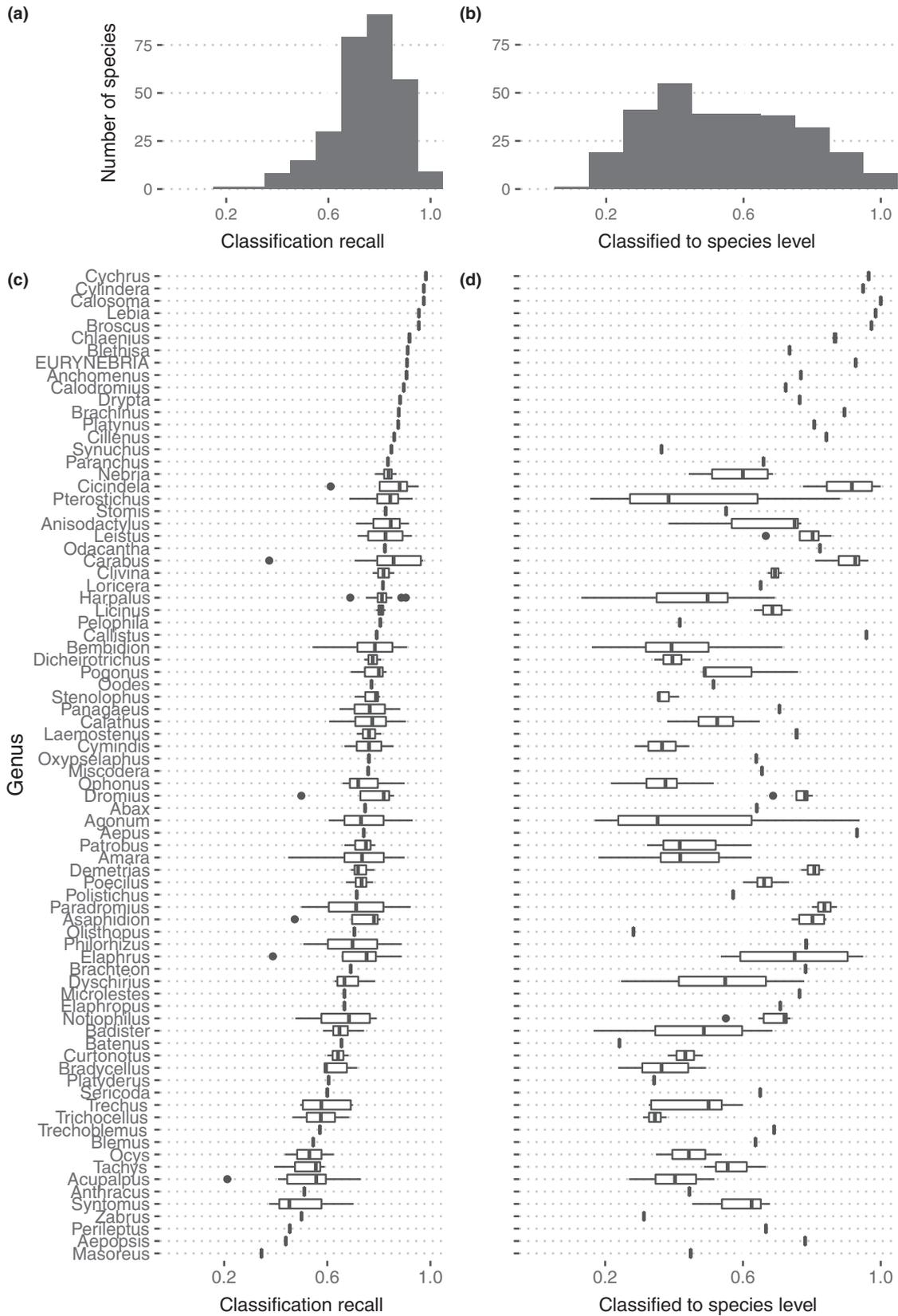


FIGURE 2 Classification performance, when setting minimum acceptable confidence threshold to 0.5. (a) Distribution and (c) genus-summary of classification recall (i.e., proportion of images of a species classified to correct taxon regardless of the predicted taxonomic level, e.g., to species, genus). (b) Distribution and (d) genus-summary of images classified to species level (i.e., proportion of images of a species classified to species level), as an indicator of classification taxonomic resolution. A large proportion of images identified to species level indicate a high taxonomic resolution, while the taxonomic resolution gradually decreases when larger proportions are identified correctly only to higher taxonomic levels (e.g., genus, tribe, or subfamily)

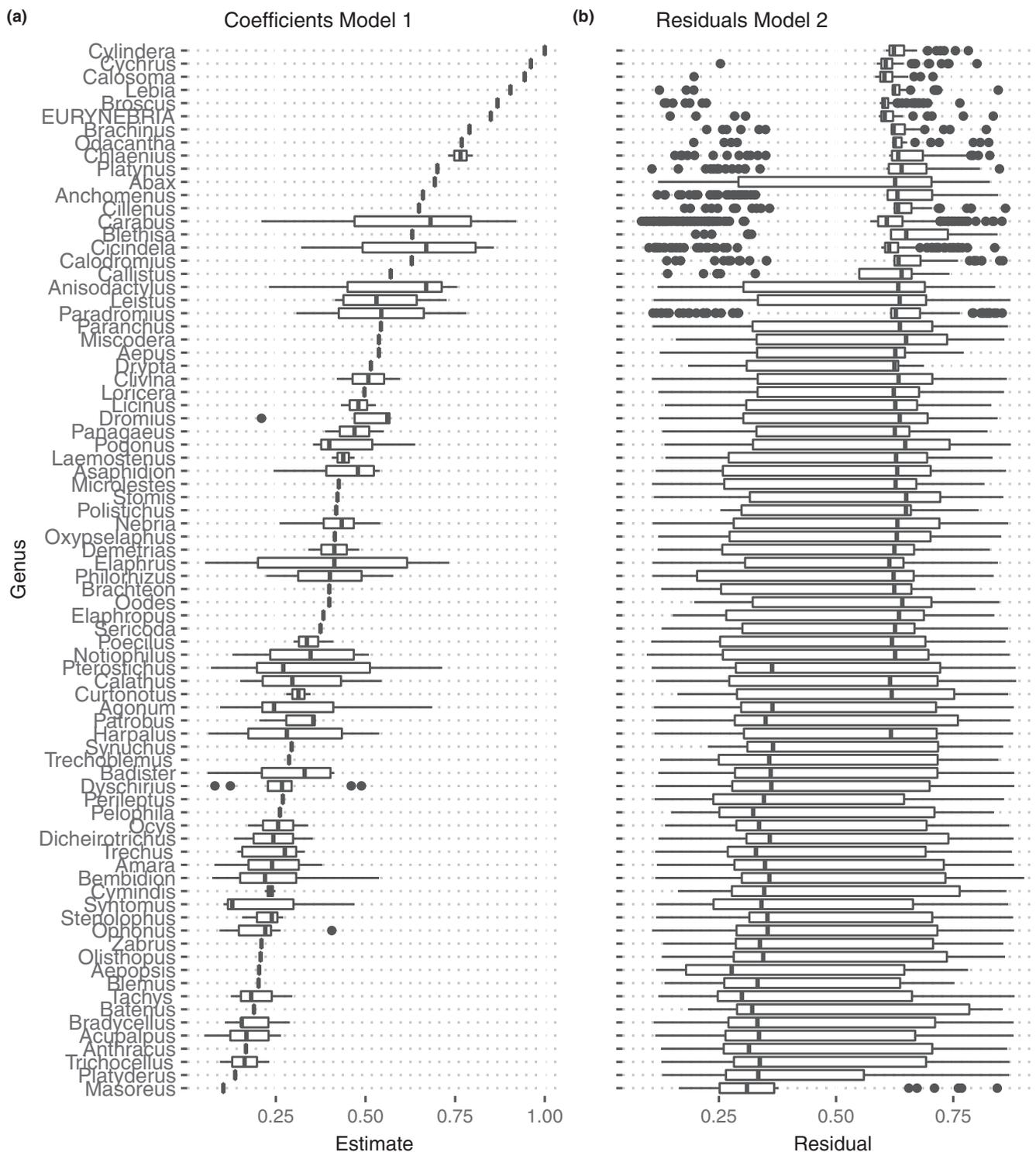


FIGURE 3 (a) Model coefficients from GLM model 1 (explanatory variable: species identity), representing species coefficients and (b) residuals from model 2 (explanatory variables: body size, top 1 value from last layer in convolutional neural network, number of species in genus, and number of training images) representing residuals for images. Species identity and number of species in genus were strongly linked, in order to keep species identity separate from other explanatory variables two models were used and residuals from model 2 compared with coefficients from model 1 compared on genus level. Higher residual values indicate that other explanatory variables than included in model 2, explain more of the variation

correctly classified images, 7,362 were correct at the species level, while the remaining 7,155 were classified correctly with less taxonomic resolution than species level. Most species had some images

predicted with a varying taxonomic resolution. Very few species had all images classified to a specific taxonomic resolution level (Figure 2b).

Parameter	Lower CI (2.5%)	Estimate	Upper CI (97.5%)
(Intercept)	0.0639	0.0713	0.0793
Top 1 value	0.986	0.988	0.990
Body size	0.616	0.659	0.701
Number of species	0.497	0.498	0.498
Number of training images	0.500	0.500	0.500

Note: Explanatory variables included the top1 value in the output layer, body size in megapixels, number of species in the same genus as the ground truth species, number of training images of the ground truth species.

Without setting a threshold, ten species had a classification recall of 90.0% or greater (Table 2; Figure 1a). With the 0.5 confidence threshold, six species had a classification recall of 90.0% or greater; however, the number of false predictions (false positives) was reduced for all ten species (i.e., increasing the recall; Table 2). With the 0.5 confidence threshold, 27 species had more than 85.0% of their test images classified at the species level, while seven species had less than 20.0% images classified at the species level (Figure S3). Genera with many species and including species which are traditionally hard to identify such as *Bembidion*, *Agonum*, *Amara*, *Harpalus*, and *Pterostichus* had a mean proportion of images classified to species level per species in the range 41.7%–45.5% (Figure S3). Inside these genera, the mean minimum and mean maximum proportion of test images classified to species level was 16.0% and 77.0%, indicating a high variability inside some genera (Figure S3).

Images classified to correct species were explained by top 1 value from the last layer in the convolutional neural network, body size, number of species within the same genus, and species identity (Figure 3; Table 3; Figure S4). The number of species within the same genus had a negative relationship with the probability of classifying to correct species while body size and the top 1 value from the last layer in CNN were positively correlated with the probability of correctly classifying the species (Table 3; Figure S4). Species identity did affect the estimate of model 1, while residuals from model 2 covaried with the estimates, suggesting that explanatory variables not included in the model could be important.

4 | DISCUSSION

Within the tested species of British Carabidae, 51.9% of the 19,164 images were classified to the correct species, when testing the model classifying to species level, and 74.9% to the correct genus, using the same trained model with genus names from ground truth and predicted species. However, the classification success for images varied significantly between species and genera, with species being everything from very difficult or very easy for the model to predict to species level. Specifically ten of 291 species had more than 90% of their images classified correctly at species level, without setting a threshold. When setting a threshold to 0.5 as minimum confidence value before decreasing taxonomic resolution, most species did,

TABLE 3 Model coefficients and 95% confidence interval from a generalized linear model (model 2), predicting if an image is classified to correct species by the convolutional neural network as binomial variable (true/false)

however, not reach taxonomic resolution at species level. The average classification recall increased from 50.7% to 75.8% using the threshold. A range of hyperparameters could have been optimized further, like learning rate and augmentation of images. This would likely have increased the overall classification recall. However, the general patterns in body size and number of species in genus would most likely remain the same. Body size of the specimens positively contributed to the models ability to classify an image. When setting a minimum threshold to the confidence level, more images were classified correctly; however, this came at the cost of losing taxonomic resolution in their prediction. In spite of the reduced taxonomic resolution, such a model can prove extremely useful in applied situations where no taxonomic information is attached beforehand, for example, reducing workload of counting, classifying on broader taxonomic levels, and creating an overview of a collection being received by a museum.

Modifying the top layer of the CNN based on the images that we extracted from the collection enabled us to distinguish among 291 classes. That is most of all species known to occur in the British Isles within ground beetles, a family belonging to one of the most species-rich orders of animals, Coleoptera with 380,000 described species (Zhang, 2011). As in other taxonomic groups, carabid beetles contain species that are morphologically only differentiated with subtle differences, which the result of this model reflected and handled to some extent by decreasing the taxonomic resolution on those image predictions, that is prediction to genus. Studies have used convolutional neural networks to classify species of a wide range of taxa, including arthropods and mammals (Norouzzadeh et al., 2018; Van Horn et al., 2017). However, this is the first to use a dataset within a well-defined geographical and taxonomic species-rich unit as well as providing information on how the postprocessing of the classification can trade-off taxonomic resolution and classification recall. As all of the images in this dataset were taken with the same fixed camera settings and distance to object, the image size could be used as a proxy for body size. Larger specimens thus have more pixels in this dataset, which is the case when scanning drawers in collections and on camera traps faced toward a ground surface, using a camera with a fixed distance to the objects. Importantly, this also suggests that images from cameras only capture a limited body size range, as images with fewer pixels are less likely to be predicted to correct species.

Comparing our results to those obtained from other convolutional neural networks, built for the specific purpose of identifying other groups of arthropods (e.g., Marques et al., 2018), there is scope for increasing both classification recall and taxonomic resolution. Either image quality, network structure, number of classes to predict, or the image recording perspective could explain the differences. When comparing precision and accuracy of dorsal perspective images of ants by Marques et al. (2018) we achieve comparable results (precision 54.7% and balanced accuracy 75.3% vs. 52.0% and 59.0%). For a range of other studies that classify arthropods to species level, the results are comparable, even though fewer species are typically used in classification (Martineau et al., 2017). van Horn et al. (2017) presented a species level trained network based on Inception ResNet v2 and 675,000 images among 5,000 species of plants and animals. Their mean accuracy level within 1,021 insect species was 74.5%, which is comparable with our balanced accuracy of 75.3% of images in our study. When critically assessing confidence values and information of taxonomy, we increased the average accuracy above the level of van Horn et al. (2017) at species level, however, losing taxonomic resolution for nearly half of the images.

Automated or semi-automated identification of insects on species or higher taxonomic levels has multiple potential applications, including in museum collections, ecological studies, and biodiversity monitoring. In museum collections, classification could identify specimens of accessions on entry into the museum and help taxonomic experts to find unusual specimens in the collections for focused taxonomic work. Classification of images enables cameras to be utilized as direct observers in ecological studies, providing detailed knowledge of species habitat preferences, activity levels, and species interactions. In monitoring, a continuous sampling of arthropod image data could also provide abilities to historically document and forecast abundances and activities of arthropods. However, all of the potential uses will only become achievable with considerable improvements of the accuracy as presented here. Proper testing and validation in applied contexts and in a broader range of taxa and habitats are crucial to achieve species-level classification.

Even though we did not find a consistent error in all species, the results indicate that CNN can be used for a variety of classification tasks with high accuracy and for some species, high taxonomic resolution. Importantly, the results indicate that habitus images are sufficient to classify images to species level, albeit not for all species. Taxonomic classification based on habitus images is needed for camera trap-based studies, where detailed images are not available. We show that assessing whether there is sufficient evidence to predict a specimen to a certain taxonomic resolution can be informed by the classification model output, through setting a confidence value threshold. Data from camera traps are possibly more complex and images from camera traps also need detection of objects, as multiple individuals may occur in the same frame. Object detection in camera traps has already been utilized for large mammals (Schneider et al., 2018), suggesting that object detection with CNN can be suitable for arthropods as well.

With the ability, from habitus images, to classify and know the classification error among arthropods including ground beetles, convolutional neural networks provide a practical tool. For ecologists, conservationists, and museum curators applied species-level classification on massive datasets can provide new opportunities for predicting the consequences of environmental changes for living organisms.

ACKNOWLEDGMENTS

Employees and volunteers at Natural History Museum Aarhus for refining boxes from full drawer scans: Birgit Balslev, Salomine Falck, Sofie Kjeldgaard. The staff and facilities at the Core Research Laboratories, Natural History Museum, London. KO acknowledges funding for this project from Innovation Fund Denmark (grant 6171-00034B), Danish Agency for Culture and Palaces under the Danish Ministry of Culture (grant FORM.2016-0025) and 15. Juni Fonden (grant 2017-N-10). JCS considers this work a contribution to his Carlsberg Foundation Semper Ardens project MegaPast2Future (grant CF16-0005) and to his VILLUM Investigator project "Biodiversity Dynamics in a Changing World" funded by VILLUM FONDEN (grant 16549). TTH acknowledges funding from his VILLUM Experiment project "Automatic Insect Detection" (grant 17523).

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

OLPH, TTH, KO, and JCS conceived the ideas; OLPH, TTH, KO, JCS, and AI designed the methodology; OLPH, BP, BG, and SD collected the data; OLPH analyzed the data; OLPH led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

A total of 63,364 images in folders corresponding to species will be released at zenodo.org under creative commons license attribution 4.0 International (<https://doi.org/10.5281/zenodo.3549369>).

ORCID

Oskar L. P. Hansen  <https://orcid.org/0000-0002-1598-5733>
 Jens-Christian Svenning  <https://orcid.org/0000-0002-3415-0862>
 Kent Olsen  <https://orcid.org/0000-0002-5624-128X>
 Steen Dupont  <https://orcid.org/0000-0001-6766-2840>
 Beulah H. Garner  <https://orcid.org/0000-0002-5229-2450>
 Alexandros Iosifidis  <https://orcid.org/0000-0003-4807-1345>
 Benjamin W. Price  <https://orcid.org/0000-0001-5497-4087>
 Toke T. Høye  <https://orcid.org/0000-0001-5387-3284>

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from <https://www.tensorflow.org/>

- Agrawal, A. A., & Inamine, H. (2018). Mechanisms behind the monarch's decline. *Science*, 360(6395), 1294–1296. <https://doi.org/10.1126/science.aat5066>
- Årje, J., Raitoharju, J., Iosifidis, A., Tirronen, V., Meissner, K., Gabbouj, M., Kärkkäinen, S. (2019). *Human experts vs. machines in taxa recognition*. Retrieved from <https://arxiv.org/pdf/1708.06899.pdf>
- Asmus, A. L., Chmura, H. E., Høye, T. T., Krause, J. S., Sweet, S. K., Perez, J. H., ... Gough, L. (2018). Shrub shading moderates the effects of weather on arthropod activity in arctic tundra. *Ecological Entomology*, 43(5), 647–655. <https://doi.org/10.1111/een.12644>
- Blagoderov, V., Kitching, I., Livermore, L., Simonsen, T., & Smith, V. (2012). No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys*, 209, 133–146. <https://doi.org/10.3897/zookeys.209.3178>
- Brown, G. R., & Matthews, I. M. (2016). A review of extensive variation in the design of pitfall traps and a proposal for a standard pitfall trap design for monitoring ground-active arthropod biodiversity. *Ecology and Evolution*, 6(12), 3953–3964. <https://doi.org/10.1002/ece3.2176>
- Caravaggi, A., Banks, P. B., Burton, A. C., Finlay, C. M. V., Haswell, P. M., Hayward, M. W., ... Wood, M. D. (2017). A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3), 109–122. <https://doi.org/10.1002/rse2.48>
- Chamberlain, S. A., & Szöcs, E. (2013). taxize: Taxonomic search and retrieval in R. *F1000Research*, 2, 191. <https://doi.org/10.12688/f1000research.2-191.v1>
- Collett, R. A., & Fisher, D. O. (2017). Time-lapse camera trapping as an alternative to pitfall trapping for estimating activity of leaf litter arthropods. *Ecology and Evolution*, 7(18), 7527–7533. <https://doi.org/10.1002/ece3.3275>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A large-scale hierarchical image database*. In IEEE Computer Vision and Pattern Recognition (CVPR). Retrieved from <http://www.image-net.org>
- Digweed, S., Currie, C., Carcamo, H., & Spence, J. (1995). Digging out the 'digging-in-effect' of pitfall traps: Influences of depletion and disturbance on catches of ground beetles (Coleoptera: Carabidae). *Pedobiologia*, 39, 561–576.
- Dolek, M., & Georgi, M. (2017). Introducing time-lapse cameras in combination with dataloggers as a new method for the field study of caterpillars and microclimate. *Journal of Insect Conservation*, 21(3), 573–579. <https://doi.org/10.1007/s10841-017-9996-9>
- Duff, A., Lott, D., Buckland, P. I., & Buckland, P. C. (2012). *Checklist of beetles of the British Isles* (173 pp.). Iver, UK: Pemberley Books.
- Dunnington, D., & Harvey, P. (2019). *exifr: EXIF Image Data in R*. Retrieved from <https://cran.r-project.org/package=exifr>
- Engel, J., Hertzog, L., Tiede, J., Wagg, C., Ebeling, A., Briesen, H., & Weisser, W. W. (2017). Pitfall trap sampling bias depends on body mass, temperature, and trap number: Insights from an individual-based model. *Ecosphere*, 8(4), e01790. <https://doi.org/10.1002/ecs2.1790>
- Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., ... De Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE*, 12(10), e0185809. <https://doi.org/10.1371/journal.pone.0185809>
- Halsall, N. B., & Wratten, S. D. (1988). The efficiency of pitfall trapping for polyphagous predatory Carabidae. *Ecological Entomology*, 13(3), 293–299. <https://doi.org/10.1111/j.1365-2311.1988.tb00359.x>
- Høye, T. T., & Forchhammer, M. C. (2008). The influence of weather conditions on the activity of high-arctic arthropods inferred from long-term observations. *BMC Ecology*, 8(1), 8. <https://doi.org/10.1186/1472-6785-8-8>
- Hudson, L. N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B. W., ... Smith, V. S. (2015). Insect: Automating the digitization of natural history collections. *PLoS ONE*, 10(11), 1–15. <https://doi.org/10.1371/journal.pone.0143402>
- Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Årje, J., ... Juhola, M. (2014). Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, 20, 1–12. <https://doi.org/10.1016/j.ecoinf.2014.01.004>
- Lister, B. C., & Garcia, A. (2018). Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National Academy of Sciences of the United States of America*, 115(44), E10397–E10406. <https://doi.org/10.1073/pnas.1722477115>
- Loboda, S., Savage, J., Buddle, C. M., Schmidt, N. M., & Høye, T. T. (2018). Declining diversity and abundance of High Arctic fly assemblages over two decades of rapid climate warming. *Ecography*, 41(2), 265–277. <https://doi.org/10.1111/ecog.02747>
- Mantle, B., LaSalle, J., & Fisher, N. (2012). Whole-drawer imaging for digital management and curation of a large entomological collection. *ZooKeys*, 209, 147–163. <https://doi.org/10.3897/zookeys.209.3169>
- Marques, A. C. R., M. Raimundo, M., B. Cavaleiro, E. M., F. P. Salles, L., Lyra, C., & J. Von Zuben, F. (2018). Ant genera identification using an ensemble of convolutional neural networks. *PLoS ONE*, 13(1), e0192011. <https://doi.org/10.1371/journal.pone.0192011>
- Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., & Venturini, G. (2017). A survey on image-based insect classification. *Pattern Recognition*, 65, 273–284. <https://doi.org/10.1016/j.patcog.2016.12.020>
- Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), E5716–E5725. <https://doi.org/10.1073/pnas.1719367115>
- Schirmel, J., Lenze, S., Katzmann, D., & Buchholz, S. (2010). Capture efficiency of pitfall traps is highly affected by sampling interval. *Entomologia Experimentalis Et Applicata*, 136, 206–210. <https://doi.org/10.1111/j.1570-7458.2010.01020.x>
- Schneider, S., Taylor, G. W., & Kremer, S. C. (2018). *Deep learning object detection methods for ecological camera trap data*. Retrieved from <http://arxiv.org/abs/1803.10842>
- Seibold, S., Gossner, M. M., Simons, N. K., Blüthgen, N., Müller, J., Ambarli, D., ... Weisser, W. W. (2019). Arthropod decline in grasslands and forests is associated with landscape-level drivers. *Nature*, 574(7780), 671–674. <https://doi.org/10.1038/s41586-019-1684-3>
- Skvarla, M. J., Larson, J. L., & Dowling, A. P. G. (2014). Pitfalls and preservatives: A review. *The Journal of the Entomological Society of Ontario*, 145, 15–43.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). Boston, MA: IEEE. <https://doi.org/10.1109/CVPR.2015.7298594>
- 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). Las Vegas, NV: IEEE. <https://doi.org/10.1109/CVPR.2016.308>
- Tensorflow, (2019). *How to retrain an image classifier for new categories*. Retrieved from https://www.tensorflow.org/hub/tutorials/image_retraining
- Van Horn, G., Mac Aodha, O., Song, Y., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2017). *The iNaturalist challenge 2017 dataset*. Retrieved from <http://arxiv.org/abs/1707.06642>
- Wagner, D. L. (2019). Insect declines in the anthropocene. *Annual Review of Entomology*, 7(1), 24. <https://doi.org/10.1146/annurev-ento-011019>
- Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, <https://doi.org/10.1111/2041-210X.13075>

- Weinstein, B. G. (2017). A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3), 533–545. <https://doi.org/10.1111/1365-2656.12780>
- Yu, X., Wang, J., Kays, R., Jansen, P. A., Wang, T., & Huang, T. (2013). Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 1(52), 1–10. <https://doi.org/10.1186/1687-5281-2013-52>
- Zaller, J. G., Kerschbaumer, G., Rizzoli, R., Tiefenbacher, A., Gruber, E., & Schedl, H. (2015). Monitoring arthropods in protected grasslands: Comparing pitfall trapping, quadrat sampling and video monitoring. *Web Ecology*, 15(1), 15–23. <https://doi.org/10.5194/we-15-15-2015>
- Zhang, Z.-Q. (Ed.). (2011). Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. In *Zootaxa* (Vol. 3148, pp. 237). Auckland, New Zealand: Magnolia Press. <http://dx.doi.org/10.11646/zootaxa.3148.1>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hansen OLP, Svenning J-C, Olsen K, et al. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecol Evol*. 2020;10:737–747. <https://doi.org/10.1002/ece3.5921>