Minimizing the Annotation Effort for Detecting Wildlife in Camera Trap Images with Active Learning

Daphne Auer[1], Paul Bodesheim[2], Christian Fiderer[3 4], Marco Heurich[5 6 7], and Joachim Denzler[8 9 10]

Abstract: Analyzing camera trap images is a challenging task due to complex scene structures at different locations, heavy occlusions, and varying sizes of animals. One particular problem is the large fraction of images only showing background scenes, which are recorded when a motion detector gets triggered by signals other than animal movements. To identify these background images automatically, an active learning approach is used to train binary classifiers with small amounts of labeled data, keeping the annotation effort of humans minimal. By training classifiers for single sites or small sets of camera traps, we follow a region-based approach and particularly focus on distinct models for daytime and nighttime images. Our approach is evaluated on camera trap images from the Bavarian Forest National Park. Comparable or even superior performances to publicly available detectors trained with millions of labeled images are achieved while requiring significantly smaller amounts of annotated training images.

Keywords: Active Learning; Wildlife Monitoring; Camera Trap Images

1 Introduction

The world experiences an accelerated loss of biodiversity that is unprecedented in the history of life [Ce15]. The major drivers of this development are habitat loss, direct persecution, climatic change, and invasive species. Therefore, the monitoring of animal populations is crucial as an early warning system and for evaluating the success of conservation measures [Se19]. Thus, national parks have installed networks of camera traps to observe

[1] Computer Vision Group, Friedrich Schiller University Jena, 07737 Jena, Germany; daphne.auer@uni-jena.de
[2] Computer Vision Group, Friedrich Schiller University Jena, 07737 Jena, Germany; paul.bodesheim@uni-jena.de
[3] Bavarian Forest National Park Germany, Visitor Management and National Park Monitoring, Freyunger Str. 2, 94481 Grafenau, Germany; christian.fiderer@npv-bw.bayern.de
[4] Wildlife Ecology and Wildlife Management, Faculty of Environment and Natural Resources, University of Freiburg, 79106 Freiburg, Germany
[5] Bavarian Forest National Park Germany, Visitor Management and National Park Monitoring, Freyunger Str. 2, 94481 Grafenau, Germany; marco.heurich@npv-bw.bayern.de
[6] Inland Norway University of Applied Science, Institute for forest and wildlife management, Campus Evenstad, NO-2480 Koppang, Norway
[7] Wildlife Ecology and Wildlife Management, Faculty of Environment and Natural Resources, University of Freiburg, 79106 Freiburg, Germany
[8] Computer Vision Group, Friedrich Schiller University Jena, 07737 Jena, Germany; joachim.denzler@uni-jena.de
[9] German Aerospace Center (DLR), Institute for Data Science, Mälzerstraße 3, Jena, Germany
[10] Michael Stifel Center Jena for Data-Driven and Simulation Science, Ernst-Abbe-Platz 2, Jena, Germany

animals like deer, roe deer, and wild boars but also smaller ones like hares and squirrels to document their appearance. This leads to thousands of images already taken in a short period, which are currently often analyzed and annotated by experts manually. Thus, an automatic species classification system would greatly enhance subsequent studies on varying animal appearances.

Furthermore, many images from camera traps do not even contain an animal but only show the background scene instead. This is often due to rapidly changing lighting or weather conditions, as well as moving plants like grasses or branches of trees caused by wind. Hence, it is not only a tedious task for humans to perform the species identification but also cumbersome to identify and filter out those background images. Our work, therefore, focuses on the automatic separation of images with and without animals to spare experts to go through uninformative data. In addition, this binary classification task serves as preprocessing step for an automatic species classification that can be applied afterwards, since the classifier does not have to recognize images showing background only anymore and can focus on the details to distinguish different animals.

To keep the manual labeling effort minimal, the task of automatically identifying background images from camera traps is tackled with an active learning approach. Starting the learning process with an initial model only trained with a small labeled training dataset, an active learning strategy automatically selects additional images that shall improve the recognition performance when annotated and incorporated in the learning process of the classifier. The aim is that only the most informative examples are chosen such that few additional annotations lead to the most significant performance gains compared to the initial model. In this work, we use the EMOC approach [FRD14] as a selection criterion in the active learning process because it has already proven to be effective in related monitoring tasks [Kä16a].

For evaluation, our approach is applied to image data from the Bavarian Forest National Park in Germany[11]. Our experiments reveal that region-specific classification models learned with only a few annotated images have the potential to outperform existing models for detecting wildlife in camera trap images, which are trained with millions of annotated images, such as the MegaDetector [BMY19] or the approach of Tabak et al. [Ta19]. Furthermore, we argue that learning separate classifiers for daytime and nighttime images is beneficial because individual models can then focus on the actual classification task compared to a single model handling both scenarios internally.

2 Related Work

Many approaches are used to automatically analyze images from camera traps, mostly adopting deep learning models for tasks like species identification and counting [VSV17, No18, Sc20, SMF20, Ta19]. Some work specifically focuses on the effect of varying sizes

[11] https://www.nationalpark-bayerischer-wald.bayern.de/english/index.htm

of the training dataset on the final species identification accuracy and showed that there is a logarithmic relationship between the number of training samples and the performance of the classifier [SMF20, Ta19]. Furthermore, the problem of class imbalance in the training data is specifically investigated with very deep network architectures [VSV17], and solutions based on transfer learning have been proposed [Wi19]. Interestingly, the authors of [Wi19] also applied their model in a live online citizen science project.

An important aspect of species identification systems is their generalization to new locations. Several studies found that most common approaches struggle when applied in previously unseen environments leading to poor performance due to the domain shift [BVHP18, Sc20]. This supports the idea of learning new, region-specific models for detection, thereby only requesting a few annotations with an active learning strategy.

As mentioned previously, we focus on the distinction between images with and without animals. This task has been tackled by [No18] with a binary classifier as well, using thousands of annotated example images. Only recently, filtering background images from camera traps in embedded systems has been proposed [Cu21]. In their study, the authors found that the performance largely depends on the number of available images for training and that neural network models with higher input resolution perform better. In contrast, our approach keeps the annotation effort minimal and uses as few labeled data as possible.

Active learning for animal detection has been previously applied [Ke19], but for drone images rather than camera trap images. However, they also propose a web-based annotation tool called AIDE that requests feedback from experts following the human-in-the-loop concept to improve a species classifier over time [KTM20]. While they use active learning for species identification in drone images, we use active learning to improve the detection of animals in camera trap images with various backgrounds and conditions.

A popular approach for analyzing camera trap images is the MegaDetector [BMY19] that has been trained on millions of example images from camera traps in North America and East Africa using the standard Faster R-CNN [Re15] as a deep object detection model. This generic detector is trained for three scenarios, i.e., it can localize animals, humans, and vehicles. Background images can be identified when even the highest detection scores for an image are below a certain threshold. Typically, a species classifier is applied for the detected animals afterwards. The MegaDetector [BMY19] is used as a baseline in our experiments.

Another detection approach that also allows for identifying background images has been proposed by Tabak et al. [Ta19]. They also provide a software package called Machine Learning for Wildlife Image Classification (MLWIC) which is used for comparison to our methods.

3 Methods

The aim of this work is to train models that separate background images from images showing an animal with only a few annotated images, which can be realized using the concept of active learning. Within the detection task, it needs to be addressed that background images and images showing an animal are very similar for a single camera trap, while the numerous backgrounds of all camera traps vary a lot, but they all belong to one common background class. That is why the task of separating images with and without animals is tackled by training region-specific models focussing on particular environments containing only a few backgrounds and perspectives.

In this work, models are trained from scratch using similar data of the region where the detector will be applied later on. Therefore, Gaussian processes are used instead of deep classifiers, since Gaussian processes already achieve good performance on very small labeled datasets, while deep classifiers require large annotated datasets. Furthermore, closed-form solutions exist for model updates of Gaussian process classifiers that are beneficial when updating the model during active learning. To also exploit recent advances in deep learning for creating meaningful image representations, the fixed CNN Inception-v3 [Sz16] with pre-trained weights from ImageNet [De09] is used as a black box feature extractor.

The Gaussian process framework and the basic idea of active learning are summarized in the following sections. Also, the sample selection strategy used for active learning is explained, namely the expected model output change (EMOC) of the Gaussian process classifier.

3.1 Gaussian Processes

The Gaussian process framework offers a probabilistic formulation for both regression and classification tasks [RW06]. The advantage is that one can directly infer uncertainties associated with the estimations of the model. These uncertainties have already been proven to be useful for active learning [Fr12, Ro17] but we will use the EMOC principle together with Gaussian processes as explained in Sect. 3.3.

The advantage of closed-form solutions for Gaussian process inference can be exploited when using label regression for classification, a common choice for visual recognition tasks [Ro17]. In this case, binary classification is achieved by taking the sign of the regression estimate when using labels $y \in \{-1, 1\}$ for the training data. Typically, labels of N training samples are collected in a vector $\boldsymbol{y}$. For the features $\boldsymbol{x}$ of the training images, one computes pairwise similarities using a kernel function κ to obtain a kernel matrix $\boldsymbol{K}$ of size $N \times N$. A common choice for two feature vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ is the Gaussian kernel with hyperparameter σ:

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right) \tag{1}$$

such that small distances in the feature space lead to high similarity values. For predicting the class of a test sample $\boldsymbol{x}^*$, it is necessary to obtain the similarities to the N training samples with the kernel function in a vector $\boldsymbol{k}_*$ of length N. This vector is then used to compute the regression estimate y^* for $\boldsymbol{x}^*$ via:

$$y^* = k_*^\mathsf{T}(\boldsymbol{K} - \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y} \tag{2}$$

with a regularization term $\sigma_n^2 \boldsymbol{I}$ that adds a small constant σ_n^2 to the main diagonal of the kernel matrix $\boldsymbol{K}$. Due to this closed-form solution, efficient updates of the model are possible with only a few algebraic operations when additional data need to be incorporated.

3.2 Active Learning

Active learning (AL) aims at improving an initial classification model trained with few data by iteratively updating the model with additional images, which are chosen using a certain strategy. In every iteration, one image is selected out of a pool of unlabeled images first. Then, it gets labeled by an expert to afterwards update the model with the additional image. This process continues until either every image is labeled, or the budget for annotation is reached. In many applications such as the analysis of camera trap images, a large amount of unlabeled data exists to choose from. The goal is to only query and annotate the meaningful images that lead to the most significant improvements of the classifier, which is guided by the above-mentioned AL strategy.

However, for experimental evaluations, a completely annotated dataset is typically used to simulate the AL pipeline in a fully automatic manner. Hence, for each image that is treated as unlabeled, the existing label will not be considered unless the image is selected by the AL strategy. Then the image label is revealed by an oracle that plays the role of the human annotator such that the labeled query image can be used to update the model. For the incorporation of additional labeled samples, we adapt a framework proposed by [Kä16b].

3.3 EMOC for Active Learning

Several strategies exist to choose an unlabeled image for additional annotation that is then used to update the model. However, this work focuses on the expected model output change (EMOC) which was already successfully applied for wildlife detection in camera trap images [Kä16a]. EMOC calculates a score for every image that tells us how much the model predictions would change if this additional image is used for extending the model.

Given an image represented by its feature vector $\boldsymbol{x}'$ and a model f, the change of the model output $\Delta f(\boldsymbol{x}')$ after updating it with $\boldsymbol{x}'$ can be calculated with the following formula:

$$\Delta f(\boldsymbol{x}') = \sum_{y' \in \mathcal{Y}} \left(\frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\boldsymbol{x} \in \mathfrak{L} \cup \mathfrak{U}} \mathcal{L}(f(\boldsymbol{x}), f'(\boldsymbol{x})) \right) p(y'|f(\boldsymbol{x}')) \tag{3}$$

Here, f' refers to the updated model and the data consist of labeled images $\mathfrak{L}$ as well as unlabeled images $\mathfrak{U}$. Furthermore, $\mathcal{L}$ is a loss function measuring the difference in model outputs $f(\boldsymbol{x})$ and $f'(\boldsymbol{x})$ before and after the model update for each image $\boldsymbol{x} \in \mathfrak{L} \cup \mathfrak{U}$. Since we do not know the label of the selected image $\boldsymbol{x}'$, we calculate the average loss (term inside large parentheses) for each possible annotation $y' \in \mathcal{Y}$ of $\boldsymbol{x}'$ and average the results with weights determined by the likelihood for the label y' of the image $\boldsymbol{x}'$.

In each iteration of the model updating process, EMOC needs to be calculated following Eq. (3) as an AL score for every unlabeled image. The unlabeled sample with the largest score is then selected for annotation and the model update. In our work, Gaussian process models are used as explained in Sect. 3.1.

4 Dataset

The methods are applied to camera trap images from the Bavarian Forest National Park, that were gathered in the years 2018 and 2019 within the project *"Neue Wege zu einem grenzüberschreitenden Rotwildmanagement in Zeiten des Klimawandels"*[12]. The collected data contains roughly 286,000 images from 100 camera traps, showing forest scenes from different perspectives day and night. Within the data, 31 native animal species could be identified by experts. The camera traps were mounted to tree trunks 60 to 70 cm above the ground and recorded a series of five images each time when motion was detected by its pyroelectric infrared sensor. However, these are still individual images taken with some seconds in between, so the camera traps did not provide continuous video clips which would have allowed for better integration of temporal context. Hence, we focus on processing the single images independently. Besides animals, there are sometimes also humans in the images, and sometimes images of one camera trap may show different forest scenes because of location changes during the operation time period.

Visibility of Animals Occlusions by vegetation, poor lighting conditions, and the distance between a camera trap and the animal mainly affect the visibility of animals in camera trap images. Sometimes, only an animal's head or leg can be seen especially if it enters the camera's field of view at the moment the image is taken. Various perspectives and backgrounds complicate the task of detecting animals in those images additionally. The animals that shall be detected vary strongly in size, ranging from larger animals like deer to smaller animals such as hares. Example images are shown in Fig. 1 which also demonstrate, that the detection can be challenging for humans as well.

Label Noise The dataset was labeled by experts with the intention to count the number of animals and define their species. As mentioned above, the camera traps took a sequence

[12] https://www.nationalpark-bayerischer-wald.bayern.de/forschung/projekte/rotwildprojekt.htm

Fig. 1: Appearing animals in camera trap images from the Bavarian Forest National Park. Already for humans, the separation of images with and without animals can be a difficult task in several situations.

of five images in a short time interval when detecting motion. Thus, the same animal(s) can be seen during a sequence and therefore the sequence needed only one annotation. Unfortunately, this label cannot be mapped to all five frames, since an animal often does not appear in the last frames anymore, leading to wrong annotations. Instead, the label is assigned to the first frame only, which reduces the dataset to about 20% of all images. Thereby, it is not guaranteed that this assignment is always correct, which explains the label noise within the data. This is why the following results need to be interpreted with caution.

Note that the following labeling is used: an image without animals is treated as a background image, including those showing humans only. This may challenge the classifier later on since all living animals belong to one class but humans need to be explicitly recognized to combine them in a class with background scene images.

Class Imbalance The ratio between images with and without animals is quite unbalanced since 68.6% of the images in the dataset do not contain an animal. The imbalance between the two classes also appears in most of the single camera trap subsets. Therefore, balanced test sets are used for evaluation, i.e., they contain the same number of testing images for each class. Of course, the class imbalance is also reflected in the set of images that the AL approach uses for drawing queries, but a reasonable selection strategy might be able to also choose selective images from the underrepresented class.

5 Experiments

For the task of separating images with and without animals, the performance of general detectors is compared with the performance of our models, which use active learning (AL) to minimize human annotation efforts. The main idea is to train one detector for a single camera trap or for a small subset of camera traps following a region-specific approach, which focuses on the corresponding scene conditions. Thereby, the number of different backgrounds is heavily reduced for a single subset, such that a model does not need to handle many backgrounds and can therefore focus on the small differences between images with and without animals. Nevertheless, the appearance of a camera trap's background still varies regarding weather and lighting conditions, as well as for different seasons.

In Sect. 5.3, experiments show that active learning with EMOC leads to competitive performance on the given task and dataset. Afterwards, those models are compared to the MegaDetector [BMY19] and the MLWIC package [Ta19] in Sect. 5.3. In Sect. 5.4 the focus lies on camera trap stations, for which these existing tools perform poorly, and it is shown that the AL approach can achieve better separations for these special cases due to well-calibrated models with a very small amount of image annotations. The dataset split for evaluating the different methods is outlined in Sect. 5.1, and the experimental setup is described in Sect. 5.1.

5.1 Dataset Preparation for Performance Comparisons

All experiments are conducted on the dataset from the Bavarian Forest National Park. As mentioned before regarding the region-specific approach, the dataset is split into small subsets, as described in Sect. 5.1.1. It also needs to be considered that each camera trap recorded a different number of images, and different ratios of class imbalance exist among the different stations, which is described in Sect. 4. In order to allow for an unbiased comparison of different approaches with respect to the class imbalance, balanced test sets of a certain size are defined in Sect. 5.1.2. Finally, camera traps that took only very few images have to be treated differently (see Sect. 5.1.3).

5.1.1 Separating Day and Night

First, the dataset is split into 100 subsets, each belonging to a single camera trap, since this work focuses on the region-specific approach. In addition, each of the subsets is separated into colored images taken during the day and grayscale images mostly taken at night, where flash is used to illuminate the scenes. The automatic separation of colored and grayscale images is easy since the corresponding color channels have characteristic values. Thus, one daytime subset and one nighttime subset are obtained for each camera trap, containing images that are illuminated either naturally or artificially by the flash. To now observe the effect of this separation on the model performance, the results for two approaches will be compared: one approach treats the daytime and nighttime subsets of a camera trap independently and two separate models are trained (dt-model and nt-model), while the other approach ignores the split and one model is trained for both daytime and nighttime images (d&nt-model). Given this distinction, a comparable setup with respect to the test sets needs to be created for the three scenarios (dt, nt, d&nt).

5.1.2 Uniform Data Splits for Balanced Test Sets

To define a uniform and consistent rule for choosing test and training data for each of the three scenarios (dt, nt, and d&nt), the variety within the numerous stations needs to be considered.

Therefore, a balanced test set for each subset is created, i.e., every test set contains the same number of images with and without animals to avoid a class imbalance in the test set. This prevents misleading accuracy values that are based on a good accuracy for the prevalent class in the data set but rather neglect the performance for the less representative one.

dt- and nt-Models After splitting the data for each camera trap into subsets for day and night, the size of the subsets ranges from only 7 images up to 3,428. Especially the class imbalance, i.e., the ratio of the two classes, differs between 97:4 to 44:42 and 6:2,291. To find general solutions for a balanced test set in each experiment, the focus is on the underrepresented class with b images. Per class, $n = floor(b/3)$ images are randomly chosen for testing, but at least five if available. For the initial training, only three images are randomly chosen per class. All remaining data serve as a pool for the AL query process.

d&nt-Models To ensure the comparability between a d&nt-model and its associated dt- and nt-model, the test set of the two models are merged for each station. Since their test sets are class-balanced, the test set for each d&nt-model also contains as many images with as without animals. The same is done with images that were chosen for the initial training and therefore six images per class images are used for the initial model for d&nt-models.

In general, it is also important to consider whether enough images are available to allow for a proper learning setup. Especially stations with only a few images either during day or night are problematic. Therefore, image subsets are merged, if they do not provide a reasonable amount of images, as described in the following section.

5.1.3 Merging Small Subsets

In the following experiments, the goal is to train a detector with only 156 train images for each subset. When splitting the subsets into their training set and test set, it turns out that 72 out of 100 dt-subsets, as well as 72 out of 100 nt-subsets, and 43 of 100 d&nt-subsets do not contain 156 images for training. Within each scenario those subsets having not enough images are merged, to generate a setup that can be applied to all experiments in Sect. 5.2. The new subgroups are created by randomly combining too small subsets until their training set includes at least 156 images.

As a result, 100 dt-subsets are merged to 48 dt-subsets, 100 nt-subsets are merged to 49 nt-subsets and 100 d&nt-subsets are merged to 73 d&nt-models. In each merging process, also the test sets are merged to a new one, instead of generating a new test set. Hence, the test sets of all d&nt-models still contain exactly the images of the test sets from the dt-models and nt-models, since the merging did not affect the test image selection. Note that models based on merged subsets need to handle more perspectives and are therefore more general, but training them is also more challenging.

5.2 Experimental Setup

After generating consistent subsets, the experimental setup can be applied to each subset within each scenario as described in the following. In every experiment, the model performance is analyzed regarding the initial accuracy, the random baseline, active learning with EMOC, and training with all available data. The initial accuracy results by training the model for the first time with only three images per class for dt- and nt-subsets and six images per class for d&nt-subsets, respectively. Afterwards, the AL process is simulated by querying 150 single images from the remaining subset iteratively and updating the model with the queried images. Choosing these images only randomly leads to a random baseline. This is compared to our AL strategy with the EMOC sample selection criterion. All experiments are repeated ten times, changing the images for the initial training randomly in each run. For comparison, another model does not only use 150 additional images for training but all data of the corresponding subset except the test set. Its performance reflects the possible enhancement using more data, and also indicates the influence of label noise and class imbalance on the training process.

Both the random selection and the selection made by EMOC generate an updated model in every querying step, which can be evaluated regarding its performance and saved afterwards. Thus, it may be possible to achieve higher accuracies with a model that was trained with fewer queried images. The main reason is the limited number of images per class in the pool of unlabeled images from which EMOC selects the queries. If there are only images of one class left after a certain amount of queries, then EMOC is also only able to select further images from the remaining class. This leads to an increasing class imbalance for the training set of the classification model. In the evaluation, the highest accuracy achieved during the AL process is therefore also reported in order to demonstrate the capacity of the approach. Automatically determining an optimal number of queries is an unsolved problem so far and needs further investigation.

For comparison, the MegaDetector [BMY19] and the MLWIC package [Ta19] are evaluated on the same test sets as the models of the AL approach. Thus, all results are affected by the label noise discussed in Sect. 4.

5.3 Overall Performance

The evaluation of the numerous models can be handled by inspecting single models regarding their overall classification accuracy, i.e., the number of correctly classified images divided by the number of all images in the test set. However, since every subset has individual parameters like the number of possible training images for the learning process or the ratio between the two classes, it is statistically more interesting to observe the performance on the whole dataset in the first place. This will be done for all models within the scenarios day, night, and day&night. Since the test sets of d&nt-models consist of the test sets of dt-models and the test sets of nt-models, the comparability of the accuracies is given.

	Our active learning approach						Baselines methods	
		Highest accuracy		Final accuracy			MegaDetector	MLWIC
	InitAcc	Random	EMOC	Random	EMOC	TrainAll	[BMY19]	[Ta19]
Only daytime images, dt-models	67.32% ±1.44%	84.62% ±0.42%	85.53% ±0.47%	79.62% ±0.34%	81.24% ±0.43%	81.76%	82.61%	67.84%
Only nighttime images, nt-models	65.25% ±1.37%	81.00% ±0.42%	81.46% ±0.43%	75.94% ±0.47%	77.01% ±0.59%	76.63%	78.15%	61.16%
All test images, dt-models + nt-models	66.25% ±0.66%	82.75% ±0.23%	83.43% ±0.33%	77.72% ±0.18%	79.06% ±0.33%	79.11%	80.31%	64.39%
All test images, d&nt-models	66.95% ±0.83%	80.76% ±0.51%	81.50% ±0.36%	76.03% ±0.66%	77.46% ±0.38%	78.85%	80.31%	64.39%
Enhancement due to daytime-nighttime separation	-0.70%	+1.99%	+1.93%	+1.69%	+1.60%	+0.26%	-	-

Tab. 1: Per-image accuracies averaged over the test images of all region subsets showing the differences between the combination of dt- and nt-models, the d&nt-models, and the baseline methods. Accuracies from the AL experiments are presented as follows: the initial accuracies obtained from a model trained on a small initial training set (InitAcc), the highest accuracy achieved during the AL process (Highest accuracy), and the final accuracy after 150 queries (Final accuracy). For comparison, the accuracies of the models learned with all training images are also reported (TrainAll). As baseline methods, the MegaDetector [BMY19] and the MLWIC package [Ta19] have been evaluated on the same test data.

Evaluation Methods Two evaluation methods are used to analyze the overall performance. One method is to treat all test sets in a scenario as a single, huge test set and evaluate the predicted labels for the corresponding images, which will be referred to as per-image accuracy. This accuracy reflects how many test images are classified correctly, which equals the probability that the prediction for a single image is correct. Another method is to evaluate how well the models perform on average for a specific region, i.e., for a single camera trap station or a subset of stations obtained from the merging of small stations. The resulting values are referred to as per-region accuracies in the following and can be interpreted as the average percentage of regions for which the predictions are correct.

For obtaining the per-region accuracies, every region subset gets the same weight when computing average values. In contrast, the per-image accuracy reflects a weighted mean using the sizes of the individual test sets as weights. As defined in Sect. 5.1.2, the size of a test set indicates the availability of the less represented class. Training a model for a binary decision with a few exemplary images for this class often leads to biased models with bad performance. By considering the per-image accuracy, the performance of models that have been trained with strongly unbalanced data is less significant. Hence, mainly the per-image accuracies are discussed since the influence of class imbalance is mitigated by considering all test images as one huge test set for the evaluation.

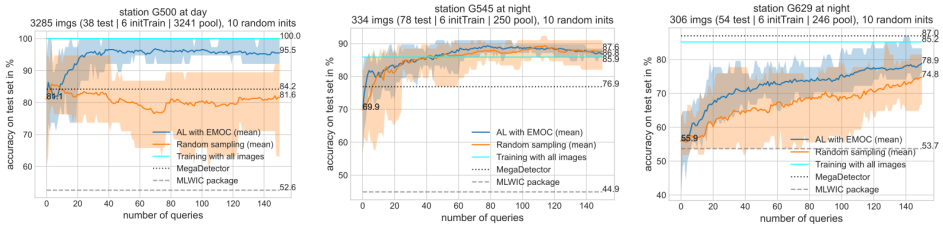
	Our active learning approach						Baselines methods	
		Highest accuracy		Final accuracy			MegaDetector	MLWIC
	InitAcc	Random	EMOC	Random	EMOC	TrainAll	[BMY19]	[Ta19]
Only daytime images, dt-models	68.08% ±1.57%	83.16% ±0.50%	84.23% ±0.38%	76.90% ±0.63%	78.17% ±0.46%	78.51%	84.74%	67.53%
Only nighttime images, nt-models	64.98% ±1.10%	80.09% ±0.60%	80.54% ±0.44%	73.93% ±0.48%	75.03% ±0.41%	74.91%	79.67%	59.61%
All test images, dt-models + nt-models	66.51% ±0.78%	81.61% ±0.25%	82.37% ±0.28%	75.40% ±0.30%	76.58% ±0.18%	76.69%	82.20%	63.53%
All test images, d&nt-models	66.94% ±0.96%	79.65% ±0.43%	80.36% ±0.32%	74.05% ±0.54%	75.40% ±0.44%	76.70%	81.71%	63.58%
Enhancement due to daytime-nighttime separation	-0.43%	+1.96%	+2.01%	+1.35%	+1.18%	-0.01%	-	-

Tab. 2: Per-region accuracies averaged over the different region subsets. Organization is the same as in Table 1, see corresponding caption explanations of the individual columns.

Separation of Daytime and Nighttime Images First, the impact of learning separate models for daytime and nighttime images is investigated. The per-image accuracies reported in Table 1 show that dt-models (first row) perform better than nt-models (second row). Although the baseline methods do not explicitly distinguish between daytime and nighttime images, they also perform better on daytime images. Thus, the distinction between images with and without animals seems to be more challenging for nighttime images. When evaluating the predictions of the individual dt-models and nt-models across all images of the test sets (third row), a clear improvement of the accuracies can be seen compared to training single d&nt-models that are able to handle both daytime and nighttime images (fourth row). The differences are indicated in the fifth row of Table 1.

It is important to note that the test sets used for the evaluations in the third and fourth row are identical, which allows for a fair comparison. This is also the reason why the per-image accuracies of the baseline methods are the same in the third and fourth row since the same baseline models are evaluated on the same test set in both lines, while for the AL experiments different models are evaluated. However, the per-region accuracies also differ for the baseline methods as shown in Table 2, because the average is taken across different region subsets of varying size but with the same weight. For the per-region accuracies, similar behavior of the individual dt- and nt-models can be observed as for the per-image accuracies, although the differences in the fifth row are less pronounced.

Active Learning Results for the AL experiments with the proposed EMOC strategy can be observed in the left parts of Tables 1 and 2. The corresponding columns are listed in order of the number of training images that have been used for learning the models. In the first column (InitAcc), the initial accuracies are reported, which denote the model performances



(a) EMOC reaches a high accuracy very fast.

(b) EMOC and random sampling perform the same way.

(c) Accuracies did not saturate after 150 queries.

Fig. 2: Visualizations of the active learning process for three example models. Training with a random sample selection is shown in orange, the learning process with EMOC as active learning strategy is shown in blue. Each line represents the mean accuracy over ten runs.

when using three training images per class for dt- and nt-models, or six training images per class for d&nt-images. Accuracies from the AL approach are shown in the following columns. While the final accuracies belong to models trained with 150 additional images, the highest accuracies are often achieved with fewer queried samples due to the reasons mentioned in Sect. 5.2. As can be seen in both tables, the final accuracies of EMOC are often very similar to the accuracies achieved by models trained with all images except those of the test set (TrainAll). Hence, it is not necessary to annotate all camera trap images for training the models in order to achieve good performance, up to 156 labeled images per region are often sufficient. Incorporating more annotated images might lead to learning biased models due to class imbalance in the training data.

EMOC vs. Random Sampling The proposed EMOC strategy is compared with a random sample selection regarding the final accuracy and the highest accuracy reached during the training process. One observation in Tables 1 and 2 is that active learning with EMOC leads to training better models in all cases compared to random sampling of images. This is consistent with the findings of previous work [Kä16a, FRD14] and shows that EMOC automatically selects meaningful images. The two different selection strategies can also be compared based on the visualizations for the AL process of three example models in Fig. 2. The optimal learning process is shown in Fig. 2a, where the pool of images is more than 20 times larger than the number of queried images. Applying the EMOC criterion results in a steep ascent in performance since it chooses meaningful images out of that large pool, in contrast to a random selection. Sometimes, there is no decisive difference between the two sampling methods like in Fig. 2b. At least, it shows that the performance for all initial runs using EMOC results in similar final accuracies, while sampling randomly often leads to more different final accuracies in each run. Some models for merged subsets, which include images from many perspectives, need more than 150 queries to reach the saturation of the training process, as demonstrated in Fig. 2c. Nevertheless, models trained with images selected by EMOC perform better than those using randomly selected images.

Comparison with MLWIC package One baseline method is provided by Tabak et al. in the MLWIC package [Ta19]. From the accuracies in Tables 1 and 2, a superior performance of the AL models can be clearly observed. Interestingly, even the initial models that are only trained with either three images per class (dt- and nt-models) or six images per class (d&nt-models) often already achieve higher accuracies, and selecting additional samples via AL even enlarges this gap. This comparison shows that applying out-of-the-box methods to an own dataset does not work in general and often leads to poor classification performance, even if these existing methods are trained with thousands of example images. The reasons are the different characteristics of the datasets that are used for training and in the application, with varying challenges and scene conditions. In contrast, better models can already be obtained using a handful of annotated images from the corresponding application domain or region, thus requiring only minimal annotation efforts.

Comparison with MegaDetector The MegaDetector [BMY19] performs better than the method of Tabak et al. [Ta19] in all scenarios regarding both the per-image accuracies (Table 1) and the per-region accuracies (Table 2). For the latter, images of more regions are classified correctly by the MegaDetector than by the models trained with our AL approach, on average. This is not surprising, since the MegaDetector has been trained with millions of annotated images. In contrast, the AL models only use about 150 to 160 annotated training images. Given this large difference in the amount of training data, the difference in per-region performance is rather low. Even more encouraging, the per-image accuracies obtained from models trained with EMOC are comparable to those of the MegaDetector when individual daytime and nighttime models are considered for the AL approach (first three rows in Table 1). Concerning the highest accuracies achieved by one of our models during the AL process, they almost identical accuracies as the MegaDetector in the aforementioned scenarios. This shows the potential for minimizing the annotation efforts for humans and keep high classification performances using an AL strategy with a clever selection of images.

5.4 Investigating Selected Region Subsets

The comparison of the AL approach with the MegaDetector [BMY19] in the previous section has shown that models trained with AL are able to achieve comparable average accuracies across all region subsets while requiring significantly less annotation effort. However, we do not state that our approach is better suited for all possible regions and camera trap scenes since it would be difficult to compete with an existing model trained on millions of images in general. In contrast, we argue that existing methods might struggle in particular regions when encountering certain scene conditions or properties of individual camera traps. Thus, models trained with AL and a small amount of additionally labeled images are one possible solution to cope with difficult or unique situations.

Therefore, ten difficult region subsets and ten easy region subsets are selected for each scenario (only daytime images, only nighttime images, both types of images) to compare

	EMOC Highest accuracy	EMOC Final accuracy	MegaDetector [BMY19]
Ten difficult region subsets w.r.t.			
Only daytime images, dt-models (564 test images | *16 camera traps*)	81.67% | *78.71%*	80.53% | *76.86%*	68.62% | *71.25%*
Only nighttime images, nt-models (481 test images | *13 camera traps*)	77.28% | *72.89%*	75.41% | *70.75%*	59.88% | *54.90%*
Both daytime and nighttime images, d&nt-models (673 test images | *10 camera traps*)	70.76% | *70.32%*	69.05% | *67.68%*	63.60% | *62.66%*
Ten easy region subsets w.r.t.			
Only daytime images, dt-models (386 test images | *17 camera traps*)	83.19% | *79.98%*	81.01% | *76.83%*	93.01% | *94.10%*
Only nighttime images, nt-models (439 test images | *17 camera traps*)	83.01% | *76.86%*	81.09% | *75.00%*	92.71% | *92.39%*
Both daytime and nighttime images, d&nt-models (543 test images | *11 camera traps*)	81.33% | *80.19%*	79.65% | *77.83%*	93.00% | *93.85%*

Tab. 3: Per-image accuracies *resp. per-region accuracies* for selected region subsets. According to the performance of the MegaDetector [BMY19], ten difficult and ten easy region subsets were chosen.

the performance of the MegaDetector with the performance of the AL approach. Here, the terms "difficult" and "easy" reference the level of difficulty for the MegaDetector as a baseline method, and do not indicate the complexity factor of the underlying task itself. The per-image accuracies and per-region accuracies are shown in Table 3.

As expected, the models trained with only a few annotated images and AL do not reach the high accuracies of the MegaDetector for the easy subsets, which are the ones where the MegaDetector performs best. These subsets contain images of situations that can be well represented by considering large amounts of images to model the still complex distribution of common scene conditions. While a large amount of training data helps to model the most likely scene conditions, the derived models such as the MegaDetector cannot capture the rare special cases in the tails of the overall data distribution, since these are underrepresented and require special attention. Thus, when looking at the region subsets that are most difficult for the MegaDetector, it can be obtained that the models trained with AL are beneficial and lead to better class separations. The reason is that these models are specifically trained for the special cases of certain region subsets using annotated images of the corresponding application domain, in contrast to generic tools like the MegaDetector. As a main result, the experiments show that one can not simply apply existing tools like the MegaDetector to an individual dataset and expect a satisfying performance for the task at hand. In contrast, it turns out that spending some effort annotating few example images for an application in a particular region and training more specific models is beneficial.

6 Conclusions and Future Work

Out-of-the-box tools for animal detection in camera trap images, like the MegaDetector [BMY19] or MLWIC [Ta19], do not perform well on images of every camera trap. One reason is that these models were trained on images from other regions like North America covering different environments and scene conditions. With active learning, an efficient method was presented to train specialized models for single camera traps, reaching competitive average performance compared to general models that have been trained with millions of images. Only 156 images were needed for training a specialized model with active learning, and the annotation effort for experts can be kept minimal. The concept allows to include new situations and sites of a camera trap as well with low annotation effort.

The investigation of single camera trap images showed, that learning region-specific models leads to improved recognition accuracies especially in regions where existing general models struggle due to difficult circumstances. Therefore, also the separation of daytime and nighttime images followed by learning individual models for these scenarios led to clear enhancements of the model performances. Another observation was that having too few images of the less represented class in the training set likely leads to low model performance since the model has only images of one class left for selecting further training data at some point. To utilize more image data of the less represented class, it is helpful to have one annotation per image instead of annotation for the whole image sequence. This would also allow for a more comprehensive analysis and for exploiting more semantic information given by the additional image labels.

The separation of images with and without animals serves as a pre-processing step for the following species classification, where empty images do not need to be identified, and the species classifier can focus mainly on animal features. In further investigations on species classification, active learning will be applied as well to choose meaningful images for training. Other investigations on an early stopping strategy for active learning would be interesting to avoid learning biased models due to class imbalance in the data. It is a direct consequence of our analyses, where it was shown that the highest model accuracies are sometimes achieved with less than 150 queried images, which have been used to obtain the final accuracies in the active learning experiments. This would also reduce the annotation efforts for humans even more.

Acknowledgements

This work has been funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF, Deutschland) via the project "Development of an Automated Multisensor Station for Monitoring of Biodiversity (AMMOD) - Subproject 5: Automated Visual Monitoring and Analysis"(FKZ: 01LC1903E).

Bibliography

[BMY19] Beery, Sara; Morris, Dan; Yang, Siyu: Efficient Pipeline for Camera Trap Image Review. In: KDD Workshop on Data Mining and AI for Conservation. 2019.

[BVHP18] Beery, Sara; Van Horn, Grant; Perona, Pietro: Recognition in Terra Incognita. In: European Conference on Computer Vision (ECCV). pp. 472–489, 2018.

[Ce15] Ceballos, Gerardo; Ehrlich, Paul R.; Barnosky, Anthony D.; García, Andrés; Pringle, Robert M.; Palmer, Todd M.: Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science Advances, 1(5), 2015.

[Cu21] Cunha, Fagner; dos Santos, Eulanda M.; Barreto, Raimundo; Colonna, Juan G.: Filtering Empty Camera Trap Images in Embedded Systems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021. Accepted for publication, arxiv preprint: https://arxiv.org/abs/2104.08859.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Fei-Fei, Li: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255, 2009.

[Fr12] Freytag, Alexander; Rodner, Erik; Bodesheim, Paul; Denzler, Joachim: Rapid Uncertainty Computation with Gaussian Processes and Histogram Intersection Kernels. In: Asian Conference on Computer Vision (ACCV). pp. 511–524, 2012.

[FRD14] Freytag, Alexander; Rodner, Erik; Denzler, Joachim: Selecting Influential Examples: Active Learning with Expected Model Output Changes. In: European Conference on Computer Vision (ECCV). pp. 562–577, 2014.

[Kä16a] Käding, Christoph; Freytag, Alexander; Rodner, Erik; Perino, Andrea; Denzler, Joachim: Large-scale Active Learning with Approximated Expected Model Output Changes. In: German Conference on Pattern Recognition (GCPR). pp. 179–191, 2016.

[Kä16b] Käding, Christoph; Rodner, Erik; Freytag, Alexander; Denzler., Joachim: Watch, Ask, Learn, and Improve: A Lifelong Learning Cycle for Visual Recognition. In: European Symposium on Artificial Neural Networks (ESANN). pp. 381–386, 2016.

[Ke19] Kellenberger, Benjamin; Marcos, Diego; Lobry, Sylvain; Tuia, Devis: Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning. IEEE Transactions on Geoscience and Remote Sensing, 57(12):9524–9533, 2019.

[KTM20] Kellenberger, Benjamin; Tuia, Devis; Morris, Dan: AIDE: Accelerating image-based ecological surveys with interactive machine learning. Methods in Ecology and Evolution, 11:1716– 1727, 2020.

[No18] Norouzzadeh, Mohammad Sadegh; Nguyen, Anh; Kosmala, Margaret; Swanson, Alexandra; Palmer, Meredith S.; Packer, Craig; Clune, Jeff: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences, 115(25):E5716–E5725, 2018.

[Re15] Ren, Shaoqing; He, Kaiming; Girshick, Ross; Sun, Jian: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems (NIPS). 2015.

[Ro17] Rodner, Erik; Freytag, Alexander; Bodesheim, Paul; Fröhlich, Björn; Denzler, Joachim: Large-Scale Gaussian Process Inference with Generalized Histogram Intersection Kernels for Visual Recognition Tasks. International Journal of Computer Vision (IJCV), pp. 253–280, 2017.

[RW06] Rasmussen, Carl Edward; Williams, Christopher K. I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. The MIT Press, 2006.

[Sc20] Schneider, Stefan; Greenberg, Saul; Taylor, Graham W.; Kremer, Stefan C.: Three critical factors affecting automated image species recognition performance for camera traps. Ecology and Evolution, 10(7):3503–3517, 2020.

[Se19] Serrouya, Robert; Seip, Dale R.; Hervieux, Dave; McLellan, Bruce N.; McNay, R. Scott; Steenweg, Robin; Heard, Doug C.; Hebblewhite, Mark; Gillingham, Michael; Boutin, Stan: Saving endangered species using adaptive management. Proceedings of the National Academy of Sciences, 116(13):6181–6186, 2019.

[SMF20] Shahinfar, Saleh; Meek, Paul; Falzon, Greg: "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecological Informatics, 57, 2020.

[Sz16] Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon; Wojna, Zbigniew: Rethinking the Inception Architecture for Computer Vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826, 2016.

[Ta19] Tabak, Michael A.; Norouzzadeh, Mohammad S.; Wolfson, David W.; Sweeney, Steven J.; Vercauteren, Kurt C.; Snow, Nathan P.; Halseth, Joseph M.; Di Salvo, Paul A.; Lewis, Jesse S.; White, Michael D.; Teton, Ben; Beasley, James C.; Schlichting, Peter E.; Boughton, Raoul K.; Wight, Bethany; Newkirk, Eric S.; Ivan, Jacob S.; Odell, Eric A.; Brook, Ryan K.; Lukacs, Paul M.; Moeller, Anna K.; Mandeville, Elizabeth G.; Clune, Jeff; Miller, Ryan S.: Machine Learning to Classify Animal Species in Camera Trap Images: Applications in Ecology. Methods in Ecology and Evolution, 10(4):585–590, 2019.

[VSV17] Villa, Alexander Gomez; Salazar, Augusto; Vargas, Francisco: Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological Informatics, 41:24–32, 2017.

[Wi19] Willi, Marco; Pitman, Ross T.; Cardoso, Anabelle W.; Locke, Christina; Swanson, Alexandra; Boyer, Amy; Veldthuis, Marten; Fortson, Lucy: Identifying animal species in camera trap images using deep learning and citizen science. Methods in Ecology and Evolution, 10:80–91, 2019.