ELPephants: A Fine-Grained Dataset for Elephant Re-Identification

Matthias Körschens[1,2,3] Joachim Denzler[2,4]

[1]Department of Plant Biodiversity, Friedrich Schiller University Jena, Jena, Germany
[2]Computer Vision Group, Friedrich Schiller University Jena, Germany
[3]Michael Stifel Kolleg, Jena, Germany
[4]Michael Stifel Center for Data Driven and Simulation Sciences, Jena, Germany
{matthias.koerschens,joachim.denzler}@uni-jena.de

Abstract

Despite many possible applications, machine learning and computer vision approaches are very rarely utilized in biodiversity monitoring. One reason for this might be that automatic image analysis in biodiversity research often poses a unique set of challenges, some of which are not commonly found in many popular datasets. Thus, suitable image datasets are necessary for the development of appropriate algorithms tackling these challenges.

In this paper we introduce the ELPephants dataset, a re-identification dataset, which contains 276 elephant individuals in 2078 images following a long-tailed distribution. It offers many different challenges, like fine-grained differences between the individuals, inferring a new view on the elephant from only one training side, aging effects on the animals and large differences in skin color.

We also present a baseline approach, which is a system using a YOLO object detector, feature extraction of ImageNet features and discrimination using a support vector machine. This system achieves a top-1 accuracy of 56% and top-10 accuracy of 80% on the ELPephants dataset.

1. Introduction

In current times a large number of changes in nature can be observed, many of which are of anthropogenic origin. Climate change, for example, is an important factor, especially for plants, which react to it locally by changing their phenology, like time of flowering [6], or globally, by shifting their occurrence into regions with more suitable climate.

For animals, other anthropogenic factors come along, like poaching and the clearing of rainforests. This can lead to emigration of animals, or, in the worst case, extinction of whole species [7], which, in turn, can lead to a strong decline in biodiversity.

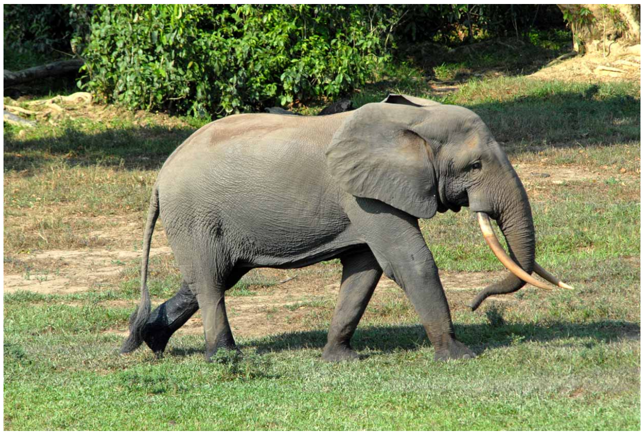

Figure 1: An example image from the dataset showing the elephant Ishmael.

To find causes of changes in phenology and abundance for plants or changes in abundance and behavior for animals, researchers have to monitor the subjects of interest over a longer period of time. For this purpose, it is beneficial to have an overview over the whole population of interest through continuous documentation.

Monitoring, identifying and documenting biological subjects can be very tedious, especially, if the number of research subjects is very large or they look very similar to each other. To alleviate the amount of work required for documenting the study subjects, automatisms are needed. In the last years, machine learning methods have become more and more popular, especially with the work of Krizhevsky *et al.* [18], following which the research efforts on convolutional neural networks steadily increased. CNNs are great methods for extracting information from images, but work best, when trained with huge numbers of images, which mostly contain classes that can easily be differentiated visually. Many popular benchmark datasets fulfill these prerequisites, for example ImageNet [28] and CIFAR-10 and CIFAR-100 [17].

In biological problems, the subjects of interest often look alike and thus the respective datasets usually are very different to the aforementioned benchmark datasets. The differences between such biological subjects are more subtle than, for example, the differences between an image of a truck and an airplane. Often a layman is not able to classify most of the subjects correctly anymore and experts are required. To automate the identification and documentation process, more specialized algorithms must be developed, which have to be designed using datasets with similar properties to the ones just described.

There are multiple larger datasets for evaluating algorithms on such fine-grained recognition tasks, like CUB200-2011 [35] and also the iNaturalist datasets [13]. However, there is another difference between these ones to most other real-world examples. This difference is the vast number of samples contained in these datasets, as they have been collected and annotated by a larger group of people.

As image acquisition and annotation can be quite expensive, common application-based datasets are much smaller and have a very heterogeneous number of images per class. Thus, while having to deal with fine-grained data, the algorithms usually also have to manage working with a small number of samples and class imbalance.

In this paper, we present the Elephant Listening Project Elephants, in short ELPephants, dataset, which can be used as a basis to develop algorithms that are able to deal with aforementioned difficulties. The animals in the images are forest elephant individuals, which have only subtle differences. We will discuss the general structure of the dataset, the usual way humans identify elephant individuals and show special characteristics of this dataset. We also discuss different issues and difficulties, which have to be dealt with when using this dataset.

Afterwards, we additionally introduce a baseline system for elephant identification, which was trained using this dataset. It can operate on one or multiple images for one identification and achieves a classification accuracy of about 56% top-1 and about 80% top-10 on the ELPephants dataset using one image. The full baseline system was presented in [15].

2. Related Work

Fine-Grained Datasets There are multiple already established fine-grained datasets, which are often used in experiments. These comprise a large range of different kinds of objects. There are datasets containing objects of technical nature, for example Stanford Cars [16] and Aircraft [23], many contain biological subjects and thus cover species of dogs (Stanford Dogs [14]) or birds, like CUB200-2011 [35] and NABirds [12]. Others deal with plants instead of animals, like Flowers102 [24], and some with more abstract objects, like food [4].

These datasets are commonly used for benchmarking new algorithms, but even though they often entail some essential characteristics of application-based biological datasets, for example classes with only fine-grained differences, they often miss one important aspect of such datasets: a small number of images. In these datasets we can still find a long-tailed distribution, but in most cases the classes still comprise quite a large number of images that can be used for training, thus nullifying the need to make the most out of a small amount of training data.

Biodiversity Datasets In the last years, multiple large-scale fine-grained classification datasets were presented in the context of biodiversity: the iNaturalist datasets [13]. These datasets comprise a large number of images, but are also highly imbalanced. The measure of imbalance, defined in [13] as the ratio of image counts of the largest class to the ones of the smallest class, is more than 29 times higher than in any of the aforementioned fine-grained datasets. Similar to the iNaturalist datasets, there are also the iWildCam sets [3], which are dealing with species (re-)identification in camera trap images.

But there are also many other datasets that deal with species identification of much smaller taxonomical groups. We can, for example, find moth datasets [27] and also datasets dealing with plants [33].

Even more complex and increasingly fine-grained are individual (re-)identification tasks. For this task we can find an Amur Tiger Dataset [19], or a number of challenges on Kaggle[1], one of which is the Humpback Whale Identification Challenge [1].

Fine-Grained Classification Fine-grained classification appears to become more relevant in the last years and there is a multitude of different approaches. Zheng *et al*. use an attention mechanism to gain more detailed information in the image [36]. Lin *et al*. developed a bilinear pooling approach [20], which has been developed further by Gao *et al*. [10] and Simon *et al*. [30, 31]. Cui *et al*. [9], in turn, have been successfully investigating domain similarity for transfer learning from a large scale dataset to a new, smaller set.

Individual Classification Individual classification has been a popular task for many years now, but mostly only in the context of human faces. Thus, especially in the age of deep learning, very powerful approaches have been developed for good face recognition [29, 25, 32].

[1] http://kaggle.com

But also in biodiversity research re-identification approaches have been developed. Loos *et al*. developed a method to re-identify great apes [22] and later improved their approach and applied it to chimpanzees [21]. Brust *et al*. similarly applied a deep learning approach for distinguishing gorilla individuals [5]. In the area of elephant identification, Ardovini *et al*. presented an approach with classical computer vision methods. They focused on mostly on the shapes and nicks of the ears of the animals to identify the elephant individuals [2].

The ELPephants dataset combines several different computer vision challenges, like fine-grained re-identification and class imbalance, with other difficulties found in biodiversity datasets, for example having to deal with differing side views of the animals, and several unique challenges, like feature occlusion through mud. In addition to this, to the best of our knowledge, the ELPephants dataset is the first elephant re-identification dataset publicly available.

3. The Dataset

The dataset introduced here comprises 2078 images of 276 elephant individuals in total with one label per image. These pictures were collected in the context of the *Elephant Listening Project*[2], in short ELP. In this project, biologists from the Cornell University in the US are conducting research on forest elephants visiting the Dzanga bai clearing in the Dzanga-Ndoki National Park in the Central African Republic.

Dzanga bai is a gathering place for many forest elephants living in the region of the national park, but most elephants only visit the clearing once every few months or even years. Having a place regularly being visited by elephants is a good opportunity for researchers to study the animals and monitor the change of their appearance and their behavior over the years. For this reason the members of the ELP have been watching this clearing for many years and documenting the different individuals during their research.

The research is being done to preserve the forest elephants in Dzanga, who are suffering from anthropogenic influence, especially poaching [34], and has been going on since the founding of the ELP in 1999. Over this time, about 4000 different elephants have been sighted and documented there. Many of these elephants were assigned names and have been recorded in photos and videos, which are taken manually by the researchers on site. We have been provided with labeled images for implementing and training a system for re-identification of the elephant individuals in an automated way. These are the images contained in this dataset.

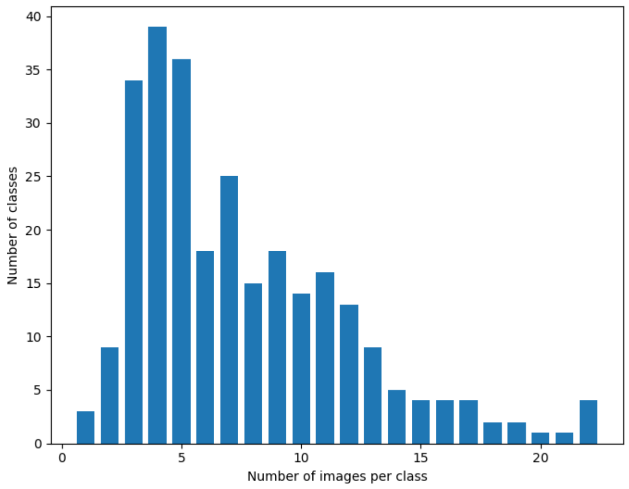

Figure 2: The number of images per class varies greatly and there are at most 22 images associated with a single class. For over half of the classes there are fewer than eight images available.

3.1. Elephant Identification Features

At first glance, elephants look very similar and are, in most cases, not easily distinguishable by laymen. But on closer inspection one can see that each elephant has characteristic features that can be used for identification. One prominent example are the tusks, which are larger for male and smaller for female elephants. The shapes of the tusks also differ greatly from one individual to another. Other important features are the body shape, and signs of fights or other natural incidents in the past. These can be, for example, scars, rips or holes in the ears, or broken tusks.

3.2. Specific Dataset Characteristics

The images have been collected over a time span of about 15 years. Because of this, we can notice an aging effect in some elephants. Young elephants get bigger and their physique changes drastically, which is especially true for males. While the physique of older elephants does not change dramatically with time, there can still be changes like new scars and broken tusks due to fights. For a human, these are the most obvious features for identifying an elephant individual, but an automatic system cannot simply rely on these features, as they might change over time.

All animals have been photographed during their stay in the clearing. Thus, there are mostly no occlusions in the images, aside from other elephants in one image.

A special characteristic of elephant images, as with many animal images, is that the look of an elephant can differ greatly depending on the perspective. The left side of the animal can look quite different from the right side, at least with respect to the more unnatural features like bro-

[2] http://www.elephantlisteningproject.org/

ken tusks, scars and rips in the ears, which are most likely not symmetrical. In addition to this, there are also images containing front views of the elephants, which differ even more from the side view for multiple reasons. For example the ears, which are often quite important for identification, might not be visible, and the shape of the tusks, which is another important characteristic, is not observable either. To tackle this problem, special mechanisms might have to be implemented.

In many biologically motivated datasets we can find a long tailed distribution, i.e. a largely imbalanced distribution of images over the classes. This is also the case for the dataset presented in this paper. The average number of images per class is about 8, but due to the large imbalance the true distribution is far-off from a uniform distribution. The actual image distribution over the classes can be seen in Figure 2. We can observe that the minimum number of images in a class is 1 and the maximum is 22. A large number of classes have only about 3-5 images, which results in 2-4 training images per individual, if the train and validation split is considered. This makes correct identification rather difficult.

3.3. Separation of Train and Validation Set

The dataset is separated into a training and a validation set by an approximate 75% stratified split. This divides the dataset into a train part of 1573 and a validation part of 505 images. Classes with only one image are only contained in the training set and classes with more than one image have at least one image in the validation set.

3.4. Difficulties and Issues in the Dataset Images

The dataset has multiple irregularities and special characteristics, which should be taken into account when developing algorithms based on it.

This dataset contains several duplicate images, about 30, and also a small number of near-duplicates, i.e. images that contain small modifications of other ones, like a change in size or brightness. These are randomly distributed over the training and validation set. A small number of incorrectly labeled elephants, i.e. label noise, can also be found in the dataset.

Several difficulties occurring in the dataset are shown in Figure 3. In image (a), an elephant is covered in mud. This results in a very bright skin and also makes it difficult to see the contours of the elephant. In addition, the large change in color might be problematic, if the algorithm has only been trained on images containing the elephant with darker skin or vice-versa. This contrast can be seen in image (d), in which the same elephant as in the other 3 topmost images is visualized. In image (b) and (c) we can see that the image is zoomed in too much. Such a zoom level hides many important features of the animal, hence making it harder to identify the elephant correctly. These four topmost images show, how large the differences in image quality can be, even for a single individual.

On the bottom left, in image (e), we see a front view of an elephant. As already stated, such images are very rare in the dataset and differ strongly from the usual side views, which makes it also harder to identify the animal correctly. Similar to this image, there are individuals, for which only one side view is contained in the training images, whereas the view of the opposite side or even the front view might be contained in the validation set. Lastly, in image (f), there are multiple elephants in the image. Here, the elephants in the background may disturb the classification of the elephant in the foreground.

4. Baseline Method

In this section, we introduce a baseline method for classifying the elephants in the proposed dataset. The original method has been described in [15] and our approach will be shown here.

4.1. The Pipeline

As a baseline method we developed a system, which is visualized in Figure 4. This system consists of multiple components that process the images sequentially within the given pipeline. In the first step an input image is processed by a YOLO network [26], which has been trained on elephant heads. In our experiments the usage of images containing only elephant heads proved to be more effective than to use the complete body of the animal. The dataset used for this training is a completely disjoint dataset from the one introduced in this paper. It contains elephant images from Flickr[3] and has been annotated manually by us with bounding boxes enclosing the ears of the elephant, the complete head and most, if not all, of the tusks and the trunk, depending on the size of these parts. This dataset consists of 1285 training and 227 test images and led to a detection precision of 92.73% and recall of 92.16% on elephant heads, with a mean average precision of 90.78% [15]. Thus the system is able to reliably detect the head of elephants.

After the detection of the head, in the case that multiple bounding boxes have been found, one box has to be selected for further processing. This is done by trying to select the most "obvious" elephant, i.e. the elephant with the largest head bounding box in conjunction with the highest confidence score provided by YOLO. This means, the selected bounding box is determined by

$$idx = \underset{i}{\operatorname{argmax}}\, A_i \cdot C_i, \tag{1}$$

with i being the index of a bounding box and A_i and C_i being the area and YOLO confidence score of each bounding

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

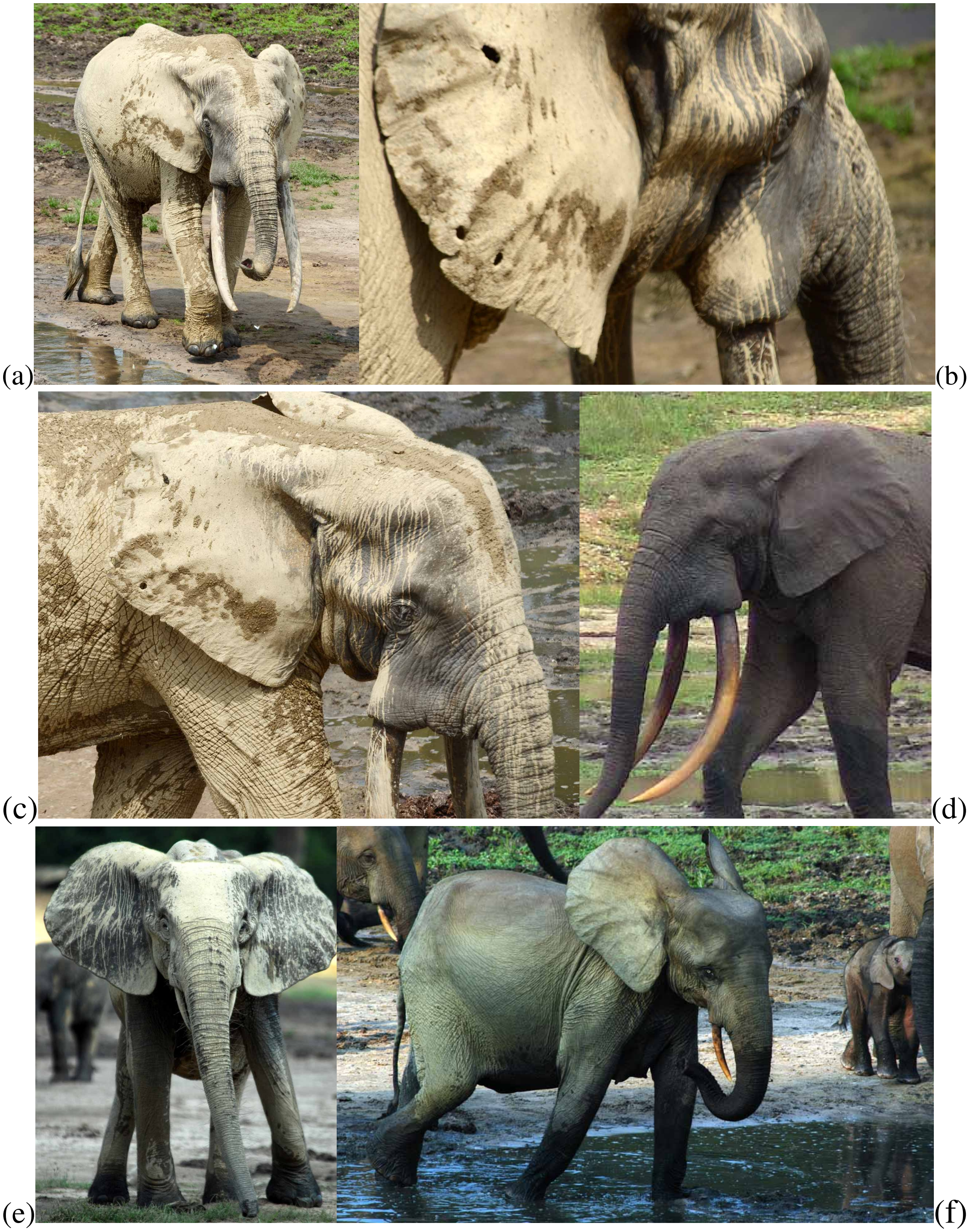

Figure 3: Possible difficulties in the dataset: (a) contours of the elephant vanish through color change caused by mud; (b) image is zoomed in too much; (c) similar to the images before: contours of the ear hardly visible and parts occluded by zoom; (d) the same elephant as is the three previous images, but with much darker skin; (e) rare front view of the elephant; (f) multiple elephants in the same image.

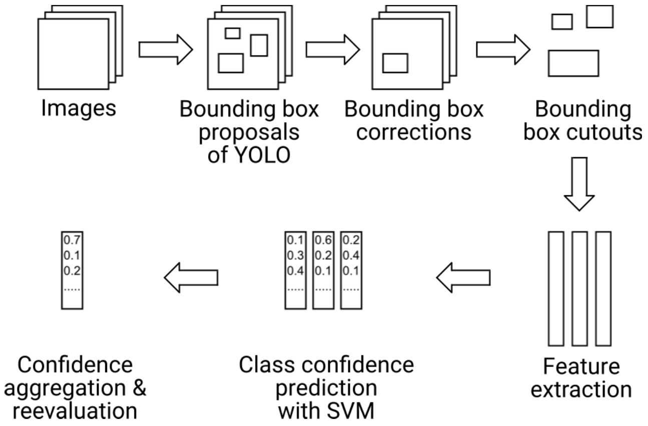

Figure 4: The full pipeline of our system from image input to the classification result. For the classification of one single individual, one or multiple images can be used. In the case of multiple images, the class confidences are aggregated to generate a joint class confidence vector.

box, respectively.

Following the selection of the most prominent bounding box, the selected one is cut out and used for feature extraction by a ResNet50 [11], which has been trained on ImageNet [28]. The extracted features are then used as input for a linear support vector machine [8], which performs a classification of the elephant.

Often, there is only a limited number of features contained in an image, especially with respect to elephant images often containing only one view of the elephant. To counter this, the system can aggregate the classification confidence scores of multiple images, which are calculated by the support vector machine. The aggregation is performed by averaging the confidence vectors to create a new classification score that takes into account multiple images and thus multiple views.

4.2. Experiments

In our experiments we discovered that feature extraction from the last activation layer in ResNet, as it is usually done, did not yield the best results. Instead, we used earlier layers in the network, i.e. the activation layers of the 13th and 14th convolutional block in the network, here denoted with "activation_40" and "activation_43", respectively.

Additionally, we added max pooling layers to increase translation invariance of the extracted features and compare multiple pooling sizes in our experiments.

The following results were generated by evaluation on the validation set.

	Top-k accuracies			
Pooling & layer	k=1	k=5	k=10	k=20
max_4, activation_40	0.508	0.706	0.770	0.823
max_5, activation_40	0.544	**0.726**	**0.800**	0.839
max_6, activation_40	**0.560**	0.716	0.788	**0.853**
max_4, activation_43	0.522	0.716	0.766	0.823
max_5, activation_43	0.546	0.708	0.770	0.833
max_6, activation_43	0.524	0.700	0.762	0.821
no pooling, activ._ 43	0.518	0.659	0.740	0.805

Table 1: Max pooling with one image using activation_40 and activation_43 features. All pooling trials were done using a network input resolution of 512×512, the trials without pooling with one of 256×256. The abbreviations max_n stand for a max pooling layer with a pooling size of $n \times n$.

	Top-k accuracies			
Pooling & layer	k=1	k=5	k=10	k=20
max_4, activation_40	0.698	0.818	0.866	0.902
max_5, activation_40	0.714	0.832	0.876	0.904
max_6, activation_40	**0.742**	**0.852**	**0.878**	0.906
max_4, activation_43	0.700	0.830	0.874	**0.908**
max_5, activation_43	0.722	0.832	0.876	0.906
max_6, activation_43	0.708	0.828	0.868	0.904
no pooling, activ._43	0.686	0.804	0.846	0.886

Table 2: Max pooling with 2 images using the layers activation_40 and activation_43. All pooling trials were done using a network input resolution of 512×512, the trials without pooling with one of 256×256. The abbreviations max_n stand for a max pooling layer with a pooling size of $n \times n$.

4.2.1 Results for One-Image-Classification

The classification results of the experiments on single test images can be seen in Table 1. We observe that the best top-1 accuracy, 56%, has been achieved by using features extracted from the activation_40 layer with 6x6 max pooling. The results using 5x5 max pooling are very similar and even outperformed 6x6 max pooling for the top-5 and top-10 results. The corresponding class-wise accuracies for the best top-1 and top-10 accuracies are 49% and 74%, respectively.

4.2.2 Results for Two-Image-Classification

Looking at the results using two test images for identifying one individual, which are shown in Table 2, we can see that the 6x6 max pooled features from activation_40 perform best with 74% top-1 accuracy and 88% top-10. The class-wise accuracies for the best top-1 and top-10 results are 59% and 79% respectively.

5. Conclusion

In this paper we introduced the new fine-grained dataset with the name ELPephants. This dataset does not only have a small number of images for comparably many classes, but also poses other additional challenges like a very small number of training images for most elephant individuals, and inference of the other side of an animal if only one side is present in the training set. In addition to these challenges, the dataset also contains multiple dataset-specific difficulties, like strong color differences of the animals, multiple elephants in one image and dealing with mud occluding crucial features.

This dataset is a valuable alternative to existing biodiversity datasets and will hopefully lead to the development of powerful systems aiding researchers in biological research disciplines.

The experiments with our proposed baseline method provide a good benchmark for other competing approaches.

Acknowledgements

We would like to express our thanks to the team of the Elephant Listening Project, especially Andrea Turkalo, Peter Wrege and Daniela Hedwig, who collected the images, made it possible to publish this dataset and provided helpful feedback.

We also thank the Carl Zeiss Foundation for the financial support within the scope of the program line "Breakthroughs".

References

[1] Humpback whale identification challenge. Kaggle Challenge, 2018. 2

[2] A. Ardovini, L. Cinque, and E. Sangineto. Identifying elephant photos by multi-curve matching. *Pattern Recognition*, 41(6):1867–1877, 2008. 3

[3] S. Beery, G. V. Horn, O. Mac Aodha, and P. Perona. The iwildcam 2018 challenge dataset. *CoRR*, abs/1904.05986, 2019. 2

[4] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 2

[5] C.-A. Brust, T. Burghardt, M. Groenenberg, C. Käding, H. Kühl, M. L. Manguette, and J. Denzler. Towards automated visual monitoring of individual gorillas in the wild. In *ICCV Workshop on Visual Wildlife Monitoring (ICCV-WS)*, pages 2820–2830, 2017. 3

[6] S. F. Bucher, P. König, A. Menzel, M. Migliavacca, J. Ewald, and C. Römermann. Traits and climate are associated with first flowering day in herbaceous species along elevational gradients. 8. 1

[7] G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, 1, 2015. 1

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 6

[9] Y. Cui, Y. Song, C. Sun, A. Howard, and S. J. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. *CoRR*, abs/1806.06193, 2018. 2

[10] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 6

[12] G. V. Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, , and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR 2015*. 2

[13] G. V. Horn, O. Mac Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. J. Belongie. The inaturalist challenge 2017 dataset. *CoRR*, abs/1707.06642, 2017. 2

[14] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011. 2

[15] M. Körschens, B. Barz, and J. Denzler. Towards automatic identification of elephants in the wild. *CoRR*, abs/1812.04418, 2018. 2, 4

[16] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 2

[17] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, institution, 2009. 2

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[19] S. Li, J. Li, W. Lin, and H. Tang. Amur tiger re-identification in the wild, 2019. 2

[20] T. Lin, A. Roy Chowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. *CoRR*, abs/1504.07889, 2015. 2

[21] A. Loos and A. Ernst. An automated chimpanzee identification system using face detection and recognition (cvpr). *EURASIP Journal on Image and Video Processing*, 2013(1):49, 2013. 3

[22] A. Loos, M. Pfitzer, and L. Aporius. Identification of great apes using face recognition. In *19th European Signal Processing Conference*, pages 922–926. IEEE, 2011. 3

[23] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2

[24] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 2

[25] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015. 2

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 779–788, 2016. 4

[27] E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. Wägele, and J. Denzler. Fine-grained recognition datasets for biodiversity analysis. *CoRR*, abs/1507.00913, 2015. 2

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 6

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, 2015. 2

[30] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *International Conference on Computer Vision (ICCV)*, pages 1143–1151, 2015. 2

[31] M. Simon, E. Rodner, T. Darell, and J. Denzler. The whole is more than its parts? from explicit to implicit pose normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2018. 2

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1701–1708, 2014. 2

[33] K. C. Tan, Y. Liu, B. Ambrose, M. Tulig, and S. Belongie. The herbarium challenge 2019 dataset, 2019. 2

[34] A. K. Turkalo, P. H. Wrege, and G. Wittemyer. Slow intrinsic growth rate in forest elephants indicates recovery from poaching will require decades. *Journal of Applied Ecology*, 54(1):153–159, 2017. 3

[35] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2

[36] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5219–5227, Oct 2017. 2