Chimpanzee Faces in the Wild: Log-Euclidean CNNs for Predicting Identities and Attributes of Primates

Alexander Freytag[1,2], Erik Rodner[1,2], Marcel Simon[1], Alexander Loos[3], Hjalmar S. Kühl[4,5], and Joachim Denzler[1,2,5]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany
[3]Fraunhofer Institute for Digital Media Technology, Germany
[4]Max Planck Institute for Evolutionary Anthropology, Germany
[5]German Centre for Integrative Biodiversity Research (iDiv), Germany

Abstract. In this paper, we investigate how to predict attributes of chimpanzees such as identity, age, age group, and gender. We build on convolutional neural networks, which lead to significantly superior results compared with previous state-of-the-art on hand-crafted recognition pipelines. In addition, we show how to further increase discrimination abilities of CNN activations by the Log-Euclidean framework on top of bilinear pooling. We finally introduce two curated datasets consisting of chimpanzee faces with detailed meta-information to stimulate further research. Our results can serve as the foundation for automated large-scale animal monitoring and analysis.

1 Introduction

In 2009, a detailed report came to the conclusion that the global biodiversity is severely threatened [31]. While the report is admiringly detailed, the assessment only represents the snapshot of a single date. However, dense information regarding the development of biodiversity over time would be highly valuable, *e.g.,* for assessing whether new political actions are required or whether previous ones have been successful. A major difficulty is the necessity of analyzing large amounts of recorded data, which often needs to be done manually. Hence, reliable quantitative insights into the status of eco systems and animal populations are difficult to obtain and expensive. In direct consequence, keeping biodiversity assessments up-to-date is rarely possible although highly needed.

While analyzing large amounts of data manually is not feasible, recording such large datasets is easily possible, *e.g.,* using camera traps [23,21]. Hence, the gap between data recording and data analysis can only be closed using reliable automated techniques. Fortunately, computer vision researchers developed a multitude of algorithms for these scenarios over the past years. Techniques of fine-grained recognition allow for visually discriminating among highly similar object categories, *e.g.,* among different birds [25], sharks [13] or flowers [15].

Fig. 1: We investigate attribute predictions for chimpanzees based on cropped faces. Results have been obtained using ground truth head regions and learned attribute predictions. *Left:* expert annotations are (Dorien-30y-Adult-Female), (Kofi-5y-Infant-Male), and (Bangolo-1y-Infant-Male). *Right:* expert annotations are (Robert-35y-Adult-Male) and (Corrie-34y-Adult-Female).

In consequence, we are in the perfect position for transferring our solutions to biologists to amplify their research.

Our first contribution is to provide such a transfer into the area of mammal investigation. More precisely, we provide an in-depth study of how to apply deep neural networks to scenarios where chimpanzees need to be analyzed. Our analysis reveals that activations of deep neural networks substantially improve recognition accuracy over established pipelines for chimpanzee identification. Moreover, they are highly useful for additional attribute prediction which allows for detailed analysis and large-scale animal monitoring. A result of our learned attribute prediction models is shown in Fig. 1.

In addition, we present how the matrix logarithm transformation can further increase discrimination abilities even on top of state-of-the-art bilinear pooling in convolutional neural networks [17]. Our technique is inspired by [30,4], where authors demonstrated its advances when using handcrafted features. The benefit in terms of recognition performance can not only be seen in our real-world application but also in a straightforward synthetic experiment that reveals the benefits of this transformation especially for fine-grained scenarios. As noise signals in low-quality images are thereby amplified as well, we found that the operation is especially helpful if the image data is of high quality.

The focus of this paper can be summarized as follows:

1. We show that deep-learned image representations significantly outperform the current state-of-the-art pipeline for chimpanzee identification,
2. We apply the LOGM-operation as post-processing on top of bilinear pooling of CNN activations and present an in-depth study of the resulting benefits, and
3. We release curated versions of the datasets presented in [19] of cropped chimpanzee faces with detailed meta information for public use.

We review related work (Section 2) and convolutional neural networks (Section 3) before introducing the matrix logarithm transform in detail (Section 4).

2 Related Work

Fine-grained Recognition Over the past decade, fine-grained recognition received increasing attention within the computer vision community due to the challenging nature of the task [3,9,8,34,25]. In contrast to classification of coarse object categories, fine-grained recognition needs to identify localized patterns, *e.g.,* striped wings or a dotted neck. A recent technique is bilinear pooling on top of CNN activations proposed by Lin et al. [17]. Furthermore, Tuzel et al. [30] and later on Carreira et al. [4] proposed the LOGM operation as post-processing of bilinear pooled handcrafted features. We combine both ideas to tackle the task of differentiating among individuals of a single species which is related to but still different from fine-grained recognition.

Identification of Human Faces Eigenfaces, one of the earliest and perhaps the most famous approach for face recognition, was presented by Turk and Pentland and is based on PCA projections of cropped face images [29]. He et al. presented Laplacianfaces, which rely on a more sophisticated projection [11]. Later on, Wright et al. reported benefits for face recognition using sparse representation models [32]. Following that line of work, Yang and Zhang improved the efficiency of sparse representation by using responses of Gabor-filters as representations [33]. Simonyan et al. transferred the idea of Fisher vector encoding to face recognition [26]. However, all of these methods rely on hand-crafted image representations which need to be optimized independently.

Recently, deep neural networks trained with millions of face images significantly improved the recognition accuracy for human faces by directly learning appropriate representations and metrics from data. One of the first networks was Deepface by Taigman et al. trained from 4M face images [27]. Even more powerful is the VGG-faces network by Parkhi et al. [22] trained from 2.6M face images. Hence, our first contribution is to adapt these models to the task of chimpanzee identification. A major difference is the comparatively small amount of training data in our application domain.

Identification of Chimpanzees To the best of our knowledge, Loos et al. [18,19] presented the only published pipeline so far for the identification of chimpanzees Inspired by results from human face recognition, a central part is the alignment of faces to guarantee that extracted visual descriptors are semantically comparable. To this end, an affine transformation is applied using facial features such as eyes and mouth and the resulting image is cropped and scaled to standard size. Aligned faces are fed into a three step pipeline which consists of feature extraction, feature space transformation, and classification. For image description, extended local ternary patterns [28] are extracted on spatially divided Gabor magnitude pictures (GMPs). Finally, the dimensionality is reduced using locality preserving projections [10] and a sparse representation classification [33] serves as classification model. Due to the implemented alignment step, the entire pipeline is restricted to near-frontal face recordings. In this work, we show how to improve accuracy by using learned image representations without the necessity of aligned face images.

3 Convolutional Neural Networks in a Nutshell

Deep (convolutional) neural networks: Computer vision systems of the last decades were commonly well-designed pipelines consisting of feature extraction, post-processing, and classification. Since all stages were developed and specified separately, this plug-and-play principle allowed for easily exchanging individual modules. In contrast, recent architectures are designed end-to-end without a clear separation of feature extraction and classification which allows for jointly optimizing all involved parameters. An example are deep neural networks which are concatenations of several processing stages f_i, $i = 1, \ldots, L$ that are tightly connected. These stages are referred to as layers and are parameterized with $\boldsymbol{\theta}_i$:

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = f_L\left(\ldots\left(f_2\left(f_1\left(\mathbf{x}_i; \boldsymbol{\theta}_1\right); \boldsymbol{\theta}_2\right)\ldots\right); \boldsymbol{\theta}_L\right) \ . \tag{1}$$

When operating on image data, location invariance of learnable patterns should be explicitly incorporated into the network layout as done in Convolutional Neural Networks (CNNs) [16]. In consequence, some layers are evaluated as convolutions between learnable filter masks and outputs of the previous layer.

Optimization: Based on collected training data $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^N$, parameter values of all layers can be estimated by jointly optimizing a single loss function:

$$\bar{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i; \boldsymbol{\theta}), y_i) + \omega(\boldsymbol{\theta}) \ . \tag{2}$$

where $\omega(\cdot)$ servers as regularizer [35]. The resulting optimization problem is usually hard and the most commonly used optimization technique is stochastic gradient descent (SGD) [2] with mini-batches [20]:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \gamma \cdot \tilde{\nabla}_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}^t; \mathcal{S}_{\text{sgd}}^t) \ . \tag{3}$$

SGD is an iterative technique where the parameter γ controls the impact of individual steps. Furthermore, the term $\tilde{\nabla}_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}^t; \mathcal{S}_{\text{sgd}}^t)$ represents the approximated gradient of the loss function with respect to the current estimate $\boldsymbol{\theta}^t$ and the currently drawn mini-batch $\mathcal{S}_{\text{sgd}}^t$:

$$\tilde{\nabla}_{\boldsymbol{\theta}} \bar{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{S}_{\text{sgd}}^t) = \frac{1}{\left|\mathcal{S}_{\text{sgd}}^t\right|} \sum_{i \in \mathcal{S}_{\text{sgd}}^t} \nabla_{\boldsymbol{\theta}} \mathcal{L}(f(\mathbf{x}_i; \boldsymbol{\theta}), y_i) + \nabla_{\boldsymbol{\theta}} \omega\left(\boldsymbol{\theta}\right) \ . \tag{4}$$

Gradients of intermediate layers can be computed using backpropagation [24].

Fine-tuning Training millions of parameters is an ill-posed problem if labeled data is rare. Fortunately, large labeled datasets exist in other application domains (*e.g.,* ImageNet [6]). The process of fine-tuning refers to using the pre-trained network weights as initialization for a novel task. Running only a limited number of optimization steps on the small data set is sufficient in practice.

4 Log-Euclidean Convolutional Neural Networks

One of the current state-of-the-art approaches on fine-grained recognition by Lin et al. [17] uses bilinear pooling to transform the outputs of convolutional layers in a CNN. Bilinear pooling has been developed by Tuzel et al. [30] and Carreira et al. [4] and computes second-order statistics of features within a spatial region. We briefly review this approach and present how the matrix logarithm on top of CNN bilinear pooling can further increase discrimination abilities.

Second-order Statistics Given the output tensor $g_{i,j,k}$ of a CNN layer with $1 \leq k \leq K$ filters, the second-order transformation is computed as pooling result over outer products of channel responses for every spatial field:

$$\boldsymbol{M} = \sum_{i,j} \boldsymbol{g}_{i,j,\cdot} \; \boldsymbol{g}_{i,j,\cdot}^T \quad . \tag{5}$$

Here, the suffix $\cdot$ denotes the vectorization of the respective component. We specified the pooling operation over spatial responses in Eq. (5) as sum-pooling, however, other pooling operations are equally possible [4].

Matrix Logarithm of Second-order Statistics The matrix $\boldsymbol{M}$ is an arbitrary symmetric positive semi-definite (PSD) matrix and therefore embedded on a Riemannian manifold but not in a (Euclidean) vector space. This implies that a function that separates the manifold into two regions (binary classification) is not just a simple hyperplane as in the Euclidean case. Whereas this is ignored in [17], it was already argued by Tuzel et al. [30] and Carreira et al. [4] that the bilinear pooling matrix should be first transformed to a vector space with proper Euclidean metric and scalar product for further processing. A straightforward option for this transformation is the matrix logarithm as used in the Log-Euclidean framework of [1], which directly maps PSD matrices to a vector space [30,4].

The matrix logarithm is computed using the eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ and performing a logarithmic transformation on the eigenvalues $\boldsymbol{D} = \mathrm{diag}(\lambda_1, ..., \lambda_K)$. Therefore, our Log-Euclidean layer performs the following operation:

$$\textsc{logm}(\boldsymbol{M}) = \boldsymbol{U} \; \mathrm{diag}(\log(\lambda_1), \ldots, \log(\lambda_K)) \; \boldsymbol{U}^T \quad . \tag{6}$$

Note that this information does not lead to any loss in information. We follow [4] and add a constant ϵ to all eigenvalues to ensure their positiveness.

Understanding the Effectiveness of the Matrix Logarithm As we will see in our experiments, the LOGM-transformation can increase accuracy in animal identification tasks. The field of topology already delivers a clear mathematical motivation. In the following, let us additionally analyze the effectiveness of LOGM from a pure machine learning point of view.

As can be seen from Eq. (6), the matrix logarithm transforms the axes length of the ellipsoid spanned by the local descriptors $\mathbf{f}_{i,j,\cdot}$. Small axes (eigenvalues) below $\alpha \approx 0.567$ get a larger absolute value. Similarly, long axes above α are shrunken with respect to their absolute value. However, the absolute value of the

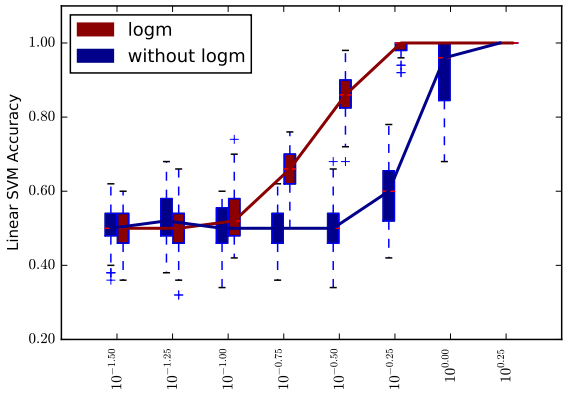

Fig. 2: Synthetic experiment comparing standard second-order pooling with LOGM-transformation of the matrices.

axes influences the impact of the axes on the matrix and also on the resulting feature vectors. Therefore, the LOGM-transformation can be seen as amplifying axes with small variances in data. Intuitively, this is ideal for identification tasks where small parts of the image are supposed to be discriminative. To verify this intuition, we performed a small synthetic experiment where we sampled matrices $\boldsymbol{M}$ for two classes as follows: the first class generates matrices by sampling 30 feature descriptors $\mathbf{g}_{i,j,\cdot}$ from a bi-variate normal distribution $\phi_1 = \mathcal{N}(\mathbf{0}, \mathrm{diag}(10^{-2}, 10))$. The second class is generated by sampling feature vectors from ϕ_1 with probability $1-p$ and from $\phi_2 = \mathcal{N}([0, \epsilon], \mathrm{diag}(10^{-2}, 10))$ with probability p. To establish a scenario that corresponds to challenging identification tasks, we use $p = 0.1$ (*i.e.,* only 10% of the feature descriptors are discriminative before bilinear pooling). We then train a linear SVM with 25 sampled matrices and evaluate the accuracy on 25 hold-out samples.

The results after 50 repetitions for various values of the distance ϵ are given in Fig. 2. As can be seen, the LOGM-transformation leads to a higher accuracy compared to standard bilinear pooling for a wide range of ϵ values (x-axis is in log-scale). In the following, we investigate the effect on real-world data.

5 Datasets for Chimpanzee Identification and Beyond

For our experiments, we assembled two datasets of cropped ape faces (denoted as C-Zoo and C-Tai). The datasets are based on previously published chimpanzee datasets by Loos and Ernst and have been extended and specifically curated for the task of attribute prediction for chimpanzee faces. We released all data as well as train-test splits at `http://www.inf-cv.uni-jena.de/chimpanzee_faces.html`. In the following, we briefly describe both datasets regarding content and quality. A detailed analysis is given in Sect. S1 of the supplementary material.

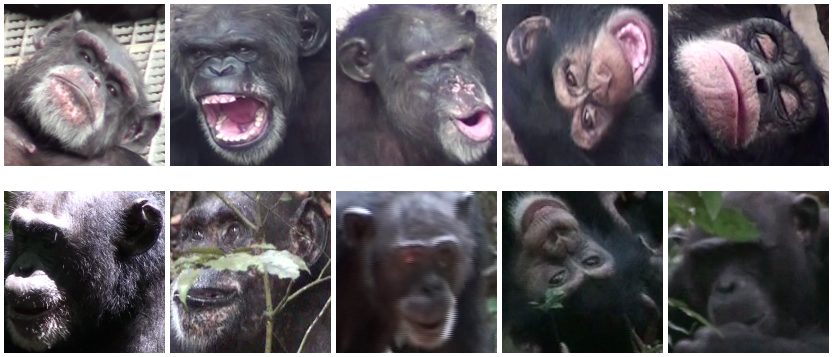

Fig. 3: Example images of the datasets C-Zoo (*top*) and C-Tai (*bottom*).

The C-Zoo Dataset Loos and Ernst introduced a chimpanzee dataset in [19] which originated from a collaboration with animal researchers in Leipzig. We build on an extension of their dataset which covers 24 individuals that have been manually labeled by experts. Provided images are of high quality, are well exposed, and are taken without strong blurring artifacts. The final C-Zoo dataset consists of 2,109 faces which are complemented by biologically meaningful keypoints (centers of eyes, mouth, and earlobes). Each individual is assigned into one out of four age groups. In addition, the gender and current age of each individual is provided as meta-information. The visual variation of contained faces is shown in the top row of Fig. 3.

The C-Tai Dataset Loos and Ernst presented a second dataset which consists of recordings of chimpanzees living in the Taï National Park in Côte d'Ivoire. The image quality differs heavily, *e.g.,* due to strong variations in illumination and distance to recorded objects. Again, we build on an extension of their data collection and obtain 5,078 chimpanzee faces which forms our second dataset. We refer to it as C-Tai and show the visual variation in the lower part of Fig. 3. In total, 78 individuals are recorded from 5 age groups. Unfortunately, the annotation quality of additional information is not as high as for the first dataset (*i.e.,* not every face is complemented with all attributes). In our evaluations, we therefor use only those 4,377 faces where identity, age, age group, and gender are provided which results in 62 different individuals.

6 Experiments

For both datasets, we were interested in the accuracy for identification of individual chimpanzees as well as for prediction of attributes age, gender, and age group. In the following, we analyze identification and gender estimation in detail for various variants of CNN codes, fine-tuning, and post processing. Due to the lack of space, we present evaluations regarding the estimation of age and age groups only in the supplementary material. The developed source code for

identification and attribute prediction including a demo-pipeline is available at `https://github.com/cvjena`.

6.1 Experiments of Chimpanzee Identification

Setup – Data For each dataset, we generate five random splits using stratified sampling with 80% for training and hold-out 20% for testing. Trained models are evaluated using averaged class-wise recognition rates to reflect the potentially imbalanced datasets (see supplementary material for dataset statistics).

Setup – Face Recognition Baselines The approach of Loos and Ernst [19] resembles the current state-of-the-art for chimpanzee recognition. We additionally follow by Parkhi et al. [22] who presented a state-of-the-art network for human face recognition. Since the network with weights trained on the Labeled Faces in the Wild (LFW) dataset [12] is publicly available, we use activations of the network for the task of chimpanzee identification (denoted with VGGFaces).

Setup – Investigated Approaches Our first question was whether CNNs for identification of human faces are better suited for the task of chimpanzee identification than other networks. Hence, we apply the Caffe BVLC reference model (denoted as BVLC AlexNet) which was originally trained for differentiating among object categories from the ImageNet challenge ILSVRC. For BVLC AlexNet and VGGFaces, we extract activations from the layers `pool5` (last layer before fully-connected layers) and `fc7` (last layer before ImageNet or LFW scores) on the cropped face regions (denoted as CNN codes). As suggested in [5], we L_2-normalize activations before passing them to the final classifier.

In addition, we were interested in the effect of post processing of CNN activations. Hence, we apply the bilinear pooling and optionally the LOGM-operation. We further increase numerical stability by normalizing the second order matrix similar to the suggestion in [17] (denoted with "+ norm", see Sect. S3 in the supplementary material). On top of extracted representations, we train linear SVMs using LibLinear [7]. The regularization parameter C is found by ten-fold cross validation in $10^{-5} \dots 10^{5}$.

Furthermore, we analyze the effect of fine-tuning with little data. To prevent over-fitting, we experiment with freezing lower layers. We follow recommendations of the Caffe toolbox and apply a weight decay of 0.0005 as well as a momentum of 0.9. On C-Zoo, we fine-tune for 2,000 iterations, whereas we conduct 8,000 iterations on C-Tai to reflect the difficulty of the dataset. The learning rate for the last layer is set to 0.001 and to 0.0001 for all remaining non-frozen layers. Training of CNNs is done using the Caffe framework [14]. During fine-tuning, we either follow the Caffe suggestions and use random crops of $227\,\text{px} \times 227\,\text{px}$ after scaling training images to $256\,\text{px} \times 256\,\text{px}$ or we directly scale images to $227\,\text{px} \times 227\,\text{px}$ (denoted as "random crops" and "no random crops"). Results for all settings are shown in Table 1.

Results – Application-specific Feature Design or Feature Learning Our first question was whether representations learned from millions of images can improve over well-designed recognition pipelines based on expert domain

Table 1: **Identification** results on C-Zoo and C-Tai. We report mean and standard deviation of avg. class-wise recognition rates (ARR) from 5 random splits (given in %).

Approach	C-Zoo	C-Tai
Baseline: state-of-the-art		
1-a) Loos and Ernst [19]	82.88 ± 01.52	64.35 ± 01.39
CNN codes + SVM		
1-b) VGGFaces `pool5`	82.73 ± 00.69	67.96 ± 01.06
1-c) VGGFaces `fc7`	66.34 ± 02.36	53.33 ± 01.04
1-d) BVLC AlexNet `pool5`	89.17 ± 01.07	$\mathbf{76.60 \pm 01.25}$
1-e) BVLC AlexNet `fc7`	81.06 ± 01.33	67.07 ± 01.58
CNN Finetuning (BVLC AlexNet)		
1-f) `fc7-fc8`, random crops	85.57 ± 05.81	51.08 ± 03.60
1-g) `fc7-fc8`, no random crops	91.89 ± 06.58	49.82 ± 03.58
1-h) `conv1-fc8`, no random crops	90.21 ± 01.66	70.22 ± 01.71
CNN codes (BVLC AlexNet) + Pooling +SVM		
1-i) `pool5` + bilinear	89.21 ± 01.59	76.13 ± 00.31
1-j) `pool5` + bilinear + norm	89.81 ± 01.25	76.22 ± 00.66
1-k) `pool5` + bilinear + norm + LOGM	$\mathbf{91.99 \pm 01.32}$	75.66 ± 00.86

knowledge. Comparing the state-of-the-art system by Loos et al. (1-a) against CNN codes (1-d) already shows a noticeable increase in accuracy. Hence, learned representations lead to clear benefits for chimpanzee identification even without further fine-tuning or post-processing.

Results – Faces or Objects Network When comparing CNN codes from VGGFaces (1-b and 1-c) and BVLC AlexNet (1-d and 1-e), we observe that the faces net is clearly outperformed. This is somewhat surprising, since we expected that the VGGFaces network could have learned typical human facial features which should also be important to distinguish between Chimpanzee faces.

Results – Fine-tuning On the C-Zoo dataset, we observe that using no random sub-crops clearly improves accuracy (1-f to 1-h). Tuning all layers slightly reduces accuracy due to overfitting to the relatively small dataset. In contrast, fine-tuning is hardly possible on the C-Tai dataset. We attribute this observation to the strong variations in pose, lighting, and occlusion which would require more data to learn a representative model.

Results – Bilinear Pooling Regarding bilinear pooling, we observe a significant increase in accuracy for the Zoo dataset when the LOGM-operation is applied (1-k). In contrast, results for the Tai dataset are not improved which we again attribute to the strong image variations in the dataset. Thereby, non-discriminative artifacts are eventually amplified by the LOGM transformation. We conclude that our transformation is well suited for identification scenarios where data is of sufficiently high quality. On C-Zoo, bilinear pooling without the LOGM transformation does not lead to a performance benefit compared to using the CNN activations directly (1-d), which we attribute to a missing vector space embedding. Finally, we observe that the LOGM-results are only marginally above results from fine-tuning on C-Zoo. However, fine-tuning requires dedicated hardware (*e.g.,* GPUs) to conduct backward passes through the network. Instead,

Table 2: **Gender estimation** results on C-Zoo and C-Tai. Results are averaged over five random splits. We report areas under ROC curves (AUC in %).

Approach	C-Zoo	C-Tai
Baseline: naive		
2-a) majority gender	50.00 ± 0.00	50.00 ± 0.00
Identification + attribute query		
2-b) using 1-k)	97.61 ± 0.94	89.60 ± 0.53
CNN codes + SVM		
2-c) VGGFaces `pool5`	94.77 ± 1.38	79.78 ± 1.94
2-d) VGGFaces `fc7`	89.32 ± 1.00	88.00 ± 0.55
2-e) BVLC AlexNet `pool5`	96.61 ± 1.07	90.49 ± 1.23
2-f) BVLC AlexNet `fc7`	95.61 ± 1.39	86.97 ± 0.62
CNN codes (BVLC AlexNet) + Pooling +SVM		
2-g) `pool5` + bilinear	97.60 ± 00.48	$\mathbf{92.97 \pm 0.42}$
2-h) `pool5` + bilinear + norm	97.81 ± 00.36	92.83 ± 0.35
2-i) `pool5` + bilinear + norm + LOGM	$\mathbf{98.16 \pm 00.35}$	90.86 ± 0.74
Cross-Dataset		
2-j) using 2-i)	70.48 ± 3.39	66.17 ± 2.73

bilinear pooling and LOGM only compute forward passes and post-processing which can efficiently be performed on low-budget standard hardware.

6.2 Evaluation of Chimpanzee Gender Estimation

Setup – Data For each dataset, we split the 2,109 and 4,377 face images by selecting 80% of each gender for training and all remaining data for model evaluation. Results are averaged over five random splits.

Setup – Baselines, Approaches, and Generalization Since Loos et al. [18,19] did not tackle attribute prediction, there is no obvious baseline for this task. Nonetheless, a naive baseline arises by predicting the majority of all genders in data ("baseline naive"). Furthermore, we can rely on the identification models of Section 6.1 and use the age of the predicted individual averaged over all its recordings during training ("Identification + attribute query"). In addition, we apply CNN codes of both networks by following the same experimental setup as in Section 6.1. Furthermore, we evaluate the effect of bilinear pooling and the LOGM-operation for the task of gender prediction. We are finally interested in the generalization abilities across datasets. Hence, we train models on LOGM-transformed features using all images from one dataset and evaluate these models on the five splits of the other dataset.

Results Results are shown in Table 2. Again, we obtain inferior results of the faces network compared to the object categorization net. Nonetheless, we observe that CNN codes on their own are already well suited for gender estimation (2-c to 2-f). The strong results are partly due to the sophisticated identification capabilities (2-a). However, bilinear pooling and the LOGM-operation can further improve results (2-g to 2-i). We finally observe that the generalization across datasets is partly possible. The clear drop in accuracy can be attributed to the different dataset characteristics (see supplementary material).

7 Conclusions

In this paper, we investigated several tasks which arise in animal monitoring for biological research. More precisely, we tackled chimpanzee identification, gender prediction, age estimation, and age group classification and provided an in-depth study of the applicability of recently popular deep neural network architectures. Furthermore, we applied the LOGM-operation as post-processing step on bilinear CNN activations which further improved accuracy when training data is sufficiently representative. Our results clearly demonstrate the effectiveness of latest vision algorithms for zoological applications, *e.g.,* with an identification accuracy of $\sim 92\%$ ARR or a gender estimation accuracy of $\sim 98\%$ AUC.

Acknowledgements The authors thank Dr. Tobias Deschner for providing the images which were used to build the C-Tai dataset, Laura Aporius and Karin Bahrke for collecting and annotating the images which were used to build the C-Zoo dataset, and the Zoo Leipzig for providing permission for image collection. The images used for creating the C-Zoo dataset were collected as part of the SAISBECO project funded by the Pact for Research and Innovation between the Max Planck Society and the Fraunhofer-Gesellschaft. Part of this research was supported by grant RO 5093/1-1 of the German Research Foundation (DFG) and by a grant from the Robert-Bosch-Stiftung.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 924–931. Springer (2006)
2. Bottou, L.: Stochastic gradient tricks. In: Neural networks: Tricks of the trade, pp. 430–445. Springer-Verlag Berlin Heidelberg (2012)
3. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Improved bird species categorization using pose normalized deep convolutional nets. In: British Machine Vision Conference (BMVC) (2014)
4. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Free-form region description with second-order pooling. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 37(6), 1177–1189 (2015)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: British Machine Vision Conference (BMVC) (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research (JMLR) 9, 1871–1874 (2008), http://www.csie.ntu.edu.tw/~cjlin/liblinear/
8. Freytag, A., Rodner, E., Darrell, T., Denzler, J.: Exemplar-specific patch features for fine-grained recognition. In: German Conference on Pattern Recognition (GCPR). pp. 144–156 (2014)

9. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2489–2496 (2014)
10. He, X., Niyog, P.: Locality preserving projections. In: Neural Information Processing Systems (NIPS). vol. 16, p. 153 (2004)
11. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 27(3), 328–340 (2005)
12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2007)
13. Hughes, B., Burghardt, T.: Automated identification of individual great white sharks from unrestricted fin imagery. In: British Machine Vision Conference (BMVC). pp. 92.1–92.14 (2015)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia. pp. 675–678 (2014)
15. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: A computer vision system for automatic plant species identification. In: European Conference on Computer Vision (ECCV). pp. 502–516 (2012)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
17. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: IEEE International Conference on Computer Vision (ICCV). pp. 1449–1457 (2015)
18. Loos, A.: Identification of great apes using gabor features and locality preserving projections. In: ACM international workshop on Multimedia analysis for ecological data. pp. 19–24. ACM (2012)
19. Loos, A., Ernst, A.: An automated chimpanzee identification system using face detection and recognition. EURASIP Journal on Image and Video Processing 2013(1), 1–17 (2013)
20. Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Le, Q.V., Ng, A.Y.: On optimization methods for deep learning. In: International Conference on Machine Learning (ICML). pp. 265–272 (2011)
21. O'Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera traps in animal ecology: methods and analyses. Springer Japan (2010)
22. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (BMVC) (2015)
23. Rowcliffe, J.M., Carbone, C.: Surveys using camera traps: are we looking to a brighter future? Animal Conservation 11(3), 185–186 (2008)
24. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature pp. 323–533 (1986)
25. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: IEEE International Conference on Computer Vision (ICCV) (2015)
26. Simonyan, K., Vedaldi, A., Zisserman, A.: Learning local feature descriptors using convex optimisation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 36(8), 1573–1585 (2014)

27. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1701–1708 (2014)
28. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing (TIP) 19(6), 1635–1650 (2010)
29. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–591 (1991)
30. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30(10), 1713–1727 (2008)
31. Vié, J.C., Hilton-Taylor, C., Stuart, S.N.: Wildlife in a changing world: an analysis of the 2008 IUCN Red List of threatened species. IUCN (2009)
32. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(2), 210–227 (2009)
33. Yang, M., Zhang, L.: Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: European Conference on Computer Vision (ECCV). pp. 448–461. Springer (2010)
34. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European Conference on Computer Vision (ECCV) (2014)
35. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301–320 (2005)