Summary: Animal Biometrics and Deep Learning

Fabian Otto*

Technische Universität Darmstadt

ABSTRACT

This summary provides an overview about the current research in Animal Biometrics, especially with regards to Deep Learning. This includes the overall processes, the filtering and preprocessing of images, architectures of networks, postprocessing, etc.

1 INTRODUCTION

The main goal in animal ecology is to observe species in their natural habitat. However, observing biodiversity can be expensive, logistically difficult and time-consuming. Many animals are rare, secretive and inhabit remote areas. Animal presence and behavior may vary over broad spatial and temporal scales, and depend on important but infrequently observed events, such as breeding, predation or mortality. Direct observation of these events can be disruptive to wildlife, and potentially dangerous to observers. To reduce cost, labor and logistics of observation, ecologists are increasingly turning to greater automation to locate, count and identify organisms in natural environments Therefore, researchers utilize camera traps, they provide a relatively cheap and easy solution to collect pictures in different areas at the same time. However, a big problem remains: The evaluation of these pictures. This summary tries to provide an overview of different approaches and techniques, which are used in Computer Vision to solve this problem. Hereby, the many main focus is to find existing research, which uses Convolutional Neural Networks (CNN), which were trained to identify individuals within one species.

2 OVERVIEW, PROBLEMS AND CLASSICAL APPROACHES

One good overview about the topic is provided by Weinstein [12]. His survey provides references to several previously conducted research projects involving Computer Vision and Animal Ecology. It also adresses other issues apart from the identification of individuals, such as counting, gender identification, species identification, description if ecological object (e.g. eggs) etc. Further, the survey demonstrates the problems current approaches have to face in order to achieve any form of classification and also presents solutions if possible.

As mentioned earlier collecting images is, due to camera traps, efficiently possible. Nonetheless, a significant issue for processing these images is their quality. It us often influenced by different illumination, shadows, weather and other image artifacts. A good overview about common problems can be found from Gómez [5] (see Fig. 1)

In order to achieve better results one suggestion from Weinstein [12] is using image metadata, such as time or location, to assist in image classification.

For individuals identification, computer vision algorithms use images of known individuals to match new images based on the similarity of phenotypic patterns. By matching the image features among images, matching algorithms score the likelihood that two images are of the same individual. For animals with unique markings, this can be a low-cost alternative to expensive trapping and tagging programs. Current state-of-the-art systems try to solve this problem with highly adapted solutions, which most of the time are not applicable for other areas/species. One example of such system is Burghardt's [6] approach to identify Great White Sharks individuals based on their fins. It detects and groups object boundaries at multiple scales into an ultrametric contour map. Afterwards, salient boundary locations are detected and used to partition region boundaries into contour sections, which are classified into fin and background classes based on shape, encoded by normals, and local appearance encoded by opponentSIFT features. Other possible areas are quite limited, it might be possible to transfer this system to other marine creates such as dolphins.

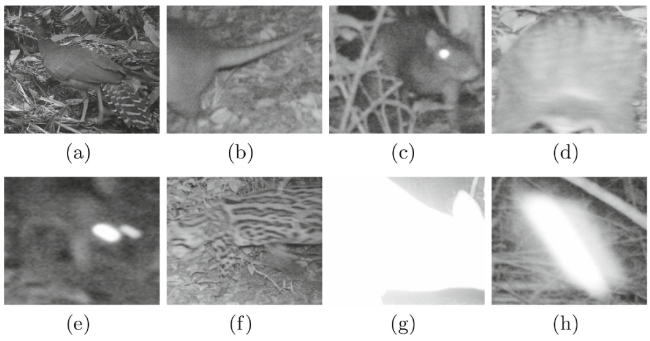

Figure 1: (a) Ideal and scarce case in camera-trap framework. (b) Partial capture of the animal. (c) Background occlusion. (d) Auto-occlusion. (e) Low resolution images. (f) Blurred animal. (g) Overexposed animal (h) Occlusion of the camera lens. [5]

3 SPECIES IDENTIFICATION

One of the first approaches for classifying species, which can be seen as similar to identifying individuals, was from Yu [13]. They reached 82% accuracy on their dataset of 7196 images and 18 classes with local and global feature extraction, which were used to train an Support Vector Machine (SVM). However, this solution still requires a lot of manual preprocessing: Removing images without animals, manually cropping all the animals from the images, and selecting only those images that captures the animals whole body.

The first completely automated approach, which used Deep Learning for species classification was established by [2]. They showed that CNNs are able to outperform the traditional Bag-of-Words technique, if enough data is provided. Before training the network, they used their own automatic segmentation method (Ensemble Video Object Cut) for cropping the animals from the images. However, the accuracy, which was achieved on their 20,000 image data set with 20 classes was only 38%. This can probably be backtracked to their relatively small network architecture, seen in Fig. 2

The work of Gómez [5] is based on the idea of clustering species to smart biologically inspired sets in order to increase the performance even for low quality as well as gray scale images. The data set of 1572 and 2597 images contains animals in the South American jungle, which present a very cluttered back ground, poor illumina-

*e-mail: fabian.otto@stud.tu-darmstadt.de

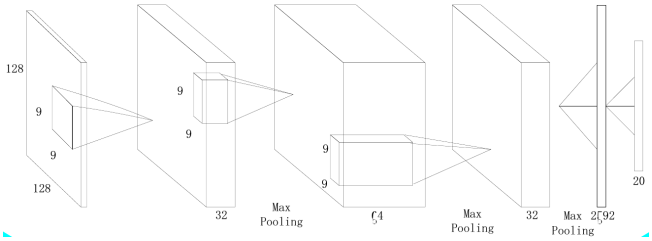

Figure 2: The structure of Chens's CNN used for species recognition. [2]

tion conditions, gray scale, and low resolution. In order to increase the trust in the system (necessary due to possible actions e.g. to save species), for each classification a confidence was computed. This confidence had to be above a threshold or the images was otherwise given to an expert. The training set in the experiments was augmented using images from the ImageNet dataset. With regards towards their architecture they utilized several state-of-the-art CNN architectures (AlexNet, VGGNet, GoogLenet and ResNet). As preprocessing the images where resized in order to fit the those. As a result they found deeper architectures work better up to ResNet-101, which was found to be the best option for the given dataset.

A similar approach is from [4], which uses the Snapshot Serengeti dataset. However they removed 22 classes that have the fewest images. Further, they split up the dataset in order to get unbalanced (rare animals less frequent), balanced (cleaned to get approx. even distribution), conditioned (animals in foreground), and segmented (manually segmented, to simulate perfect segmentation algorithm, which finds (part of) the animal) datasets, which were used during their analysis. The used dataset contains images with high as well as low quality such as in Fig. 3.

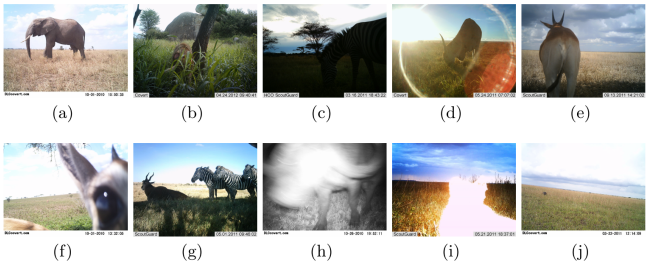

Figure 3: (a) Ideal. (b) Occlusion due to context. (c) Poor illumination. (d) Over-exposed regions. (e) Auto-occlusion. (f) Complex animal poses and unexpected images. (g) Different species in the same image. (h) Blurred. (i) Over-exposed animals. (j) Animals far way from camera. [4]

They also compared several state-of-the-art CNN architectures (AlexNet, VGGNet, GoogLenet and ResNet). The CNNs were all pre-trained with ImageNet and fine-tuned with Stochastic Gradient Descent (SGD) and back-propagation. The Learning rate and step size were reduced to avoid overfitting. They found even the deepest architectures can not deal with highly unbalanced data. But still perform better than the more shallow networks. For balanced datasets the accuracy increases to 70%. Problems are empty frames or high distance between camera and animal. The conditioned data increased accuracy by 12.9%, the segmented dataset, however, increased the conditioned result only slightly. The suggestion of the paper is to use deeper networks and skip segmentation as well as use sufficiently large datasets.

Another approach based in the Snapshot Serengeti dataset can be found from [9, 10]. Their research goal was to identify species in the Snapshot Serengeti dataset including all 3.2 million images with 48 species. Additionally, they removed images with multiple labels/multiple animals and trained CNNs for different tasks:

- animal vs. empty
- animal identification (species)

Preprocessing includes rescaling the images and normalizing all three input channels individually (z-transformation). The training is conducted by using SGD with momentum (0.9), weight decay and batch size of 128.

The best result was achieved by using VGGNet. Nonetheless, ResNet architectures show similar performance with decreases less than 1% in accuracy. In their results they also provide a comparison between pre-trained CNNs as they were used by [4] and CNNs trained from scratch. There results suggest that the features learned from the ImageNet dataset do not provide an advantage for learning animal classification. They propose to use pre-trained models on more similar tasks/input, so that Deep Learning can also be applied for data sets containing less images than Snapshot Serengeti. Further, their findings show ensembles increase the overall accuracy. One issue with this approach is low accuracy for rare classes. The paper uses weighted loss, oversampling and emphasis sampling to overcome these problems. However, only some classes show improved results with all three methods.

4 Fine-Grained Object Recognition

All the above approaches try to solve a more coarse grained problem with species identification. Even though some species might be really similar, this is definitely true for individuals of the same species.

One approach can be found from Freytag et al. [3]. Their goal is to identify identity, age, age group, and gender of chimpanzee faces in the wild without the necessity of aligned face images. As mentioned in Sect. 2 this is often done by highly specific hand-crafted recognition pipelines which are prone to noise. In their approach they define the problem as fine-grained recognition and utilize the bilinear pooling approach from [7]. Further they use matrix logarithm on top fo the CNN bilinear pooling (*LOGM*-transformation). This 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. However, the paper also suggests that low-quality images are thereby amplified as well. The overall approach was to extract features, which were found in CNNs, apply bilinear pooling and optionally *LOGM*-transformation and finally pass them to a linear SVM after L2-normalizing it. Training was done by SGD with weight decay (0.0005), momentum (0.9) and learning rate 0.001 to 0.0001. Regarding preprocessing, they found that random crops of 227 px $\times$ 227 px after scaling training images to 256 px $\times$ 256 px performed worse than directly scaling images to 227 px $\times$ 227 px. The experiments support that bilinear pooling and *LOGM*-transformation show worse results on data with more noise. Additionally, they found that pre-trained models on human faces [8] show much worse results than models trained from scratch. Which is similar to [9, 10], where pre-training on ImageNet decreased the performance. Further, feature extraction worked better right before the first fully connected layer compared to after.

This approach motivates to go in the direction of fine-grained recognition. A paper published based on fine-grained recognition is from Branson [1] which is working on bird species classification on a dataset containing 11,788 images of 200 bird species. The goal was to classify bird species with the help of pose normalization based on aligning detected keypoints to the corresponding keypoints in a prototype image. Therefore, they used annotated data, which included parts (namely: image, bounding box, body, head) and keypoints in order to learn how the pictures have to be warped (see

Fig. 4). Based on that they learn the constellation of parts via the deformable part model (DPM) nad normalize the parts accordingly.

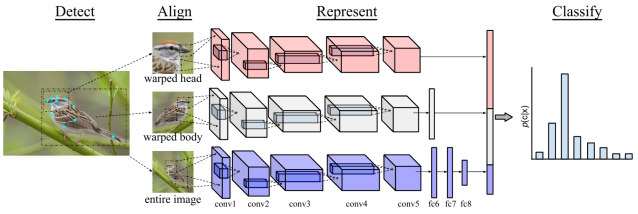

Figure 4: Branson's pipeline overview [1]

They also evaluated a pre-trained model on ImageNet and a fine-tuned version. Additionally, they used a second approach, which fine-tuned only the fully connected layer. The newly learned weights were then used for initialization during the training of the whole network. They saw significant improvements due to the use of CNN features. However, their model drops in performance if no ground truth parts are given and detected parts are used. This means, manual annotation of parts is required. Further, they found pre-training with ImageNet essential in order to achieve good results. Comparing both fine-tuning methods, the second approach worked more reliable over multiple trails.

The above mentioned approach from Lin [7] focuses more generally on fine-grained recognition. Among others the same data set as in [1] was used. They describe an approach, which has two feature extractors based on CNNs whose outputs are multiplied using the outer product at each location of the image and pooled across locations to obtain an image descriptor (see Fig. 5]). Their experiments are also different from [1] as they do not assume annotated parts or bounding boxes.

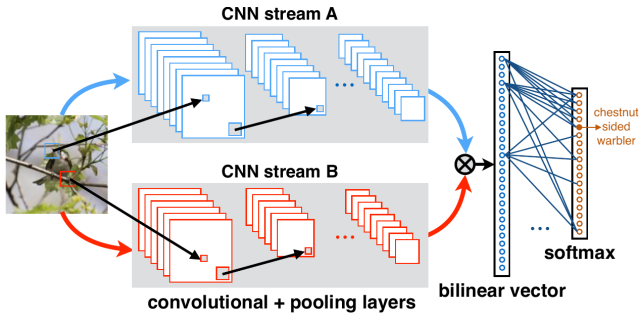

Figure 5: A bilinear CNN model for image classification. At test time an image is passed through two CNNs, A and B, and their outputs are multiplied using outer product at each location of the image and pooled to obtain the bilinear vector. This is passed through a classification layer to obtain predictions. [7]

For their experiments they used two M-Nets, two VGGNets and a combination with one of each. They pre-trained their CNNs on ImageNet and used only the convolutional layers, which were then combined with bilinear pooling and l2-normalized. For classification they used logistic regression or linear SVM, which can be replaced by multi-layer neural network for non-linear outputs. Further, they compare the results for input data with given bounding boxes and without bounding boxes. Both results where equally good with 84.1% accuracy and 85.1% respectively for the combination of M-Net and VGGNet. However, two VGGNets performed equally well. Problems, which could be found, were the bilinear CNN models that are symmetrically initialized will remain symmetric after fine-tuning. Therefore, they suggest to include Dropout or dimensionality reduction with Principal Component Analysis (PCA) at the end of one of the CNNs. Nonetheless, the performance did not increase.

A similar, but different direction is proposed by Simon [11]. Their goal is to learn model parts completely unsupervised, unlike [1]. In order to get part proposal they use pre-trained CNNs with ImageNet. Also other data sets are possible as the CNN can be pre-trained on a weakly related object dataset. The later layers in that CNN are sensitive to increasingly abstract patterns in the image. These patterns can correspond to whole objects or parts of objects, which as used as part proposals. Nonetheless, one problem is the output resolution at these layers. To solve this problem deep neural activation maps are computed, which are used consecutively in order find the part constellations. Therefore, different views of objects, i.e. a selection of part proposals are used. Shift vectors denote the offset to the common root location of an object. The goals is then to optimize the selected view, the common root, the parts proposals associated with a view and if a part is visible or not (see Fig. 6).

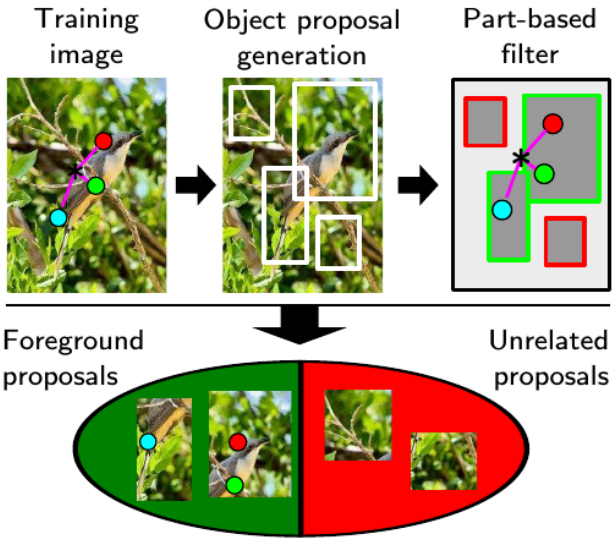

Figure 6: Overview of Simon's approach to filter object proposals for fine-tuning of CNNs [11]

As result they mainly found improvement, compared to a plain VGGNet without applying part constellations, when classifying the bird datasets. On other datasets the performance was similar.

5 Conclusions and Ideas

References

[1] S. Branson, G. V. Horn, S. J. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014.

[2] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 858–862, Oct 2014. doi: 10.1109/ICIP.2014.7025172

[3] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kuhl, and J. Denzler1. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. *GCPR: Pattern Recognition*, 9796, 2016.

[4] A. Gómez, A. Salazar, and F. V. Bonilla. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *CoRR*, abs/1603.06169, 2016.

[5] A. Gomez Villa, G. Diez, A. Salazar, and A. Diaz. Animal identification in low quality camera-trap images using very deep convolutional neural networks and confidence thresholds. vol. 10072, pp. 747–756, 12 2016.

[6] B. Hughes and T. Burghardt. Automated visual fin identification of individual great white sharks. *CoRR*, abs/1609.06323, 2016.

[7] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference on Computer Vision (ICCV)*, 2015.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 1:41.1–41.12, 01 2015.

[9] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, C. Packer, and J. Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *CoRR*, abs/1703.05830, 2017.

[10] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, C. Packer, and J. Clune. Automatically identifying wild animals in camera trap images with deep learning. *CoRR*, abs/1703.05830, 2017.

[11] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. *CoRR*, abs/1504.08289, 2015.

[12] B. G. Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545. doi: 10.1111/1365-2656.12780

[13] X. Yu, J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013(1):52, Sep 2013. doi: 10.1186/1687-5281-2013-52