Animal Identification in Low Quality Camera-Trap Images Using Very Deep Convolutional Neural Networks and Confidence Thresholds

Alexander Gomez[1], German Diez[1], Augusto Salazar[1(✉)], and Angelica Diaz[2]

[1] Grupo de Investigación SISTEMIC, Facultad de Ingeniería, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
alexander.gomezvilla@supsi.ch, augusto.salazar@udea.edu.co
[2] Instituto Alexander Von Humboldt, Calle 28A No. 15-09, Bogota D.C, Colombia

Abstract. Monitoring animals in the wild without disturbing them is possible using camera trapping framework. Automatic triggered cameras, which take a burst of images of animals in their habitat, produce great volumes of data, but often result in low image quality. This high volume data must be classified by a human expert. In this work a two step classification is proposed to get closer to an automatic and trustfully camera-trap classification system in low quality images. Very deep convolutional neural networks were used to distinguish images, firstly between birds and mammals, secondly between mammals sets. The method reached 97.5% and 90.35% in each task. An alleviation mode using a confidence threshold of automatic classification is proposed, allowing the system to reach 100% of performance traded with human work.

1 Introduction

Currently, automated camera-traps used in wildlife are small devices, fixed to a plant, rock or other structure. Camera-traps are powerful tools for wildlife scientists, whose, by using this method, can answer fundamental questions and resolve issues like: detecting rare species, delineating species, distributions, documenting predation, monitoring animal behaviour, and other vital rates [1]. Hence, it allows biologists to protect animals and their environments from extinction or man-made damage.

Camera-trapping generates a large volume of images. Therefore, it is a big challenge to process the recorded images and it is even harder, if the biologists are looking to identify all photographed species. Currently, no automatic approach is used to identify species from camera-trap images. Researchers analyse thousands or millions of photographs manually [2]. An automatic system that deals with this problem would accelerate the professionals' work, allowing them to focus on data analysis and important issues only.

Automatic classification of animal species in camera-trap images has been approached in very unrealistic or specific scenarios. A few previous works proposed solutions for this problem. Yu et al. [3] manually cropped and selected

G. Bebis et al. (Eds.): ISVC 2016, Part I, LNCS 10072, pp. 747–756, 2016.
DOI: 10.1007/978-3-319-50835-1_67

images, which contain the whole animal body. This conditioning allowed then to obtain 82% of accuracy classifying 18 animal species in their own dataset. Although Chen et al. [4] use an automatic segmentation algorithm and did not manually select images they obtained only 38.3% of accuracy. Finally Gomez et al. [5] got 88.9% in challenging scenarios without manually selecting images used high quality images with three channels.

In this work very deep convolutional neural networks are used to classify between animal species sets. Instead of classifying each image as belonging to a species a partition using a hierarchy based on biologist knowledge was used. First, all the input images are classified as birds or other. Secondly, the other cluster is classified as big mammals or small mammals, each group has several species tidied up by biologist.

Our result show that, if direct animal species classification is avoided, and the species are clustered in smart biologically inspired sets, high performance even in low quality and gray scale images is possible. Also trusting in the system prediction only when a confidence threshold is surpassed and leave hard decisions to a human expert leads to nearly perfect performance and confidence.

The rest of the paper is organized as follows. Related work is mentioned in Sect. 2. In Sect. 3 the challenges present in camera- trapping framework are described; also the methods used in the identification model are explained. Section 4 describes the experiments used to test the models. Results are presented in Sect. 5. Finally, in Sect. 6 conclusions and future work are presented.

2 Related Work

This section reviews previous approaches to identify species in camera-trap images. To the best of our knowledge there are only three previous approaches to identify animal species in camera-trap images. Sparse coding spatial pyramid matching (ScSPM) with a linear support vector machine was used by Yu et al. [3] to recognize 18 species of animals, reaching 82% of accuracy on their own dataset. As input to the ScSPM the photo-trap images were preprocessed by removing empty frames (images without animals), manually cropping all the animals from the images, and selecting only those images that capture the animals' whole body without distortions or noise. This procedure gave Yu et al. high performance but in unrealistic conditions.

A deep convolutional neural network (ConvNet) was used by Chen et al. [4] to classify 20 animal species in their own dataset. An important difference from [3] is that they use an automatic segmentation method (Ensemble Video Object Cut) for cropping the animals from the images and use this crops to train and test their system. Although this is a realistic classification scenario the classification model is too weak to capture class variability. Also the automatic segmentation algorithm gave Chen et al. a lot of empty images which give them a 38.31% of accuracy.

Very deep ConvNets were used by Gomez et al. [5] to classify 26 animal species in the Snapshot Serengeti dataset. Multiple versions of training and testing samples as unbalanced samples, empty frames, incomplete animal images,

cropped animals and objects too far from focal distance were used. A comparison using the Chen et al. dataset showed that deeper architectures outperform shallow ones in the same dataset. Although the performance of the models of Gomez et al. showed high values in the Snapshot Serengeti dataset, said model performance in Chen et al.'s dataset was less than 60% due to a high-noisy training set. This fact reveals a lack of confidence in low quality camera-trap images.

Our approach uses very deep ConvNets as [5], but is different in two main aspects. First, unlike the Snapshot Serengeti dataset, our data contain images of animals in the south american jungle, which presents a very cluttered background, also poor illumination conditions, gray scale, and low resolution. Second, we do not directly classify animal species but animal sets proposed by biologist in order to make a trade-off between classification complexity and human effort.

3 Methods

In this section different situations, present in camera-trap images, that must be overcome to make species identification automatically, are described and analysed. Also, a solution based on very deep convolutional neural networks is proposed.

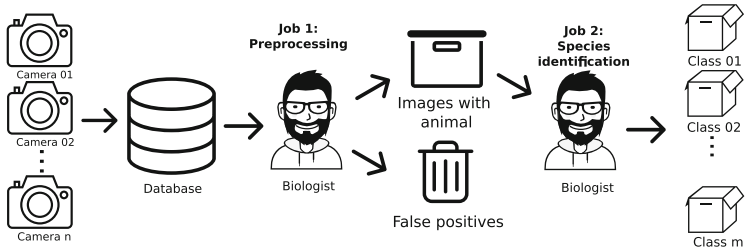

Fig. 1. Framework of data processing in camera-trap

3.1 Framework of Data Processing in Camera-Trap

The framework of data processing in camera-trap is showing in Fig. 1. All images are put in a database, which is processed in 2 ways: First (Job 1), all the images without animals (false positive) are removed. This pre-process is the most time-consuming step since typically more than 50% of the images are false positives. Then the images than contains animals are classified in species (Job 2) using the biologist knowledge to identify visually and using metadata information as hour and temperature. A few works has tried to solve Job 1. Analizing Job 2 as a computer vision problem it implies a segmentation of the animal and a classification of the segmented region. In this work only the classification part of Job 2 is solved.

The image classification problem is interpreted as an object recognition problem in which for instance a bird (see Fig. 4(b)) must be recognized. However as were said in previous works this kind of prototype images are an ideal and scarce case in camera-trap framework. Common camera-trap images are corrupted by eight possible noise conditions.

First and most common problem is partial capture of the animal (see Fig. 2(b)). This problem can confuse any automatic algorithm that learned a class appearance based in an specific feature (like head attributes or legs shape) even biologist are unable to identify some images in this condition. Background occlusion is a common feature of camera-trap images in jungle or dense forest. Unlike Snapshot Serengeti dateset where African savannah makes difficult to occlude any animal in jungle or forest is very common. Although camera-traps must be set in a clean region plants grown or animals continuously enter and exit from vegetation zones. Occlusion has the same effect of partial images and additionally introduces a lot of noise in the animal body regions.

Auto-oclussion (see Fig. 2(d)) has the same effect of partial images, complex poses that hides main features of the species are commonly found in camera-trap images. Low resolution images (see Fig. 2(e)) can be consequence of hardware selection or animal behaviour. High resolution cameras can be used but as result less cameras can be acquired. Even if the camera has high resolution animals do not always walk near the camera as is expected or the scene is too deep as African savannah. Low resolution images can make animal undistinguishable even for a human expert. Blurred animals (see Fig. 2(f)) are also consequence of hardware and animal behaviour unlikely partial images do not hide animal features but reduce the confidence of classification. Overexposed animal images (see Fig. 2(g)) occurs when the animal is too near of camera flash this effect can erase completely distinctive skin patterns of the animal body. Finally, occlusion of the camera lens (see Fig. 2(h)) by water from the environmental conditions can contaminate a large set of images and introduces partial images, occluded images and even blurred sections.

3.2 Alleviation Mode

In camera-trap framework biologist have a lot of responsibility in species classification task, since some species are very unlikely to appear on image (like jaguar in Colombian camera-traps). If one of this scarce species are omitted the whole region study is wrong, since this species are in extinction danger or gives a lot of biological information. Trust in an automatic classifier means take the bet of the automatic system will never incorrectly classify an image that have this scarce species.

A multinomial logistic regression gives a probability of membership to each class computed. When a new sample is processed if the probability of membership to each class is very similar the classifier, actually the model is more likely to assign it to an incorrect class [6]. In this situation is more secure to let a human expert to classify the sample. This process allow a trade-off between

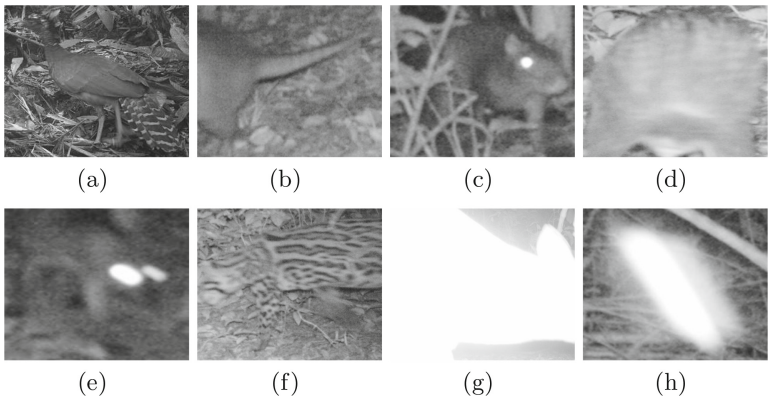

Fig. 2. (a) Ideal and scarce case in camera-trap framework. (b) Partial capture of the animal. (c) Background occlusion. (d) Auto-oclussion. (e) Low resolution images. (f) Blurred animal. (g) Overexposed animal (h) Occlusion of the camera lens.

human work and performance of the system and has been used previously in machine learning systems [7]

$$confidence = \mid log_{10}(Prob_A/Prob_B) \mid \tag{1}$$

The confidence of a classification is computed using Eq. 1. Where $Prob_A$ is the probability of membership to class one and $Prob_B$ is the probability of membership to class two. If the confidence of a classification is less that a threshold level called alleviation this classification is passed to a human expert.

3.3 Convolutional Neural Networks

Convolutional neural networks [8] consist of stacked convolutional and pooling layers ending in a fully connected layer with a feature vector as Fig. 3 shows as output. Convolutional layers generate feature maps followed by a non-linear activation function. Pooling layers provides scale invariant capacity to the extracted features. A common topology in a ConvNet consists of many sequential stacked convolutional and pooling layers that can extract discriminative features from an input image.

A transformation that maps from low level to high level features is done in a ConvNet. The first layers contain low level features (e.g., edges and orientation) and the last layers contain high level representation features, such as wrinkles, or in the case of animals, the fur details and its discriminative patterns. An important issue in ConvNets architectures is the Depth, reason why community attempts to boost topology Depth. In this work AlexNet [9], VGGNet [10], GoogLenet [11], and ResNets [12] are used in order to probe how Depth in ConvNets impacts in camera trapping species recognition.

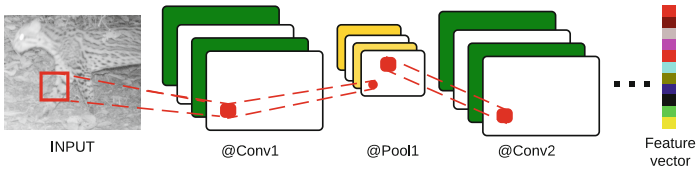

Fig. 3. Convolutional neural network

4 Experimental Framework

In this section the datasets used and the experiments carried out in this work, are described. Additionally, an explanation of implementation details (such as libraries and architecture parameters) is included.

4.1 Dataset

Our dataset is composed of 95000 images took in Colombian territory by the Von Humboldt Institute. This images are a collection of camera-trap captures taken in nine sites. The raw dataset was pre-processed by human experts cutting out the animals present on the images. Nevertheless a high percentage of the total amount of images not contain animals, but vegetation or useless information instead of identifiable animal parts. The used dataset was a subset of the original group of 95000 images where the original images were segmented only on the animal containing parts. The segmented dataset was split in two experiments. The total number of images for each experiment is 1572 for Experiment 1 and 2597 for Experiment 2. This images were split in test and train sets, as Table 2 shows. Notice the unbalanced nature of the training set in both experiments.

4.2 Experiments

The ideal classification task is distinguishing species, however, like previous works with similar images showed, the performance with these low-quality images is far from acceptable. In this work two experiments proposed by biologist are done. First, a separation between birds and other animals is done, since camera-trap framework is not designed for birds, however, birds appear on scene and activate camera sensors. This filter is the first step to alleviate the experts' work. Once all the birds are removed from the set the second experiment classify between two mammals sets. Further details of theses sets are shown in Table 1.

Although this separation is no per-species it gives a lot of useful information to the biologists. The second filter is an easier task for the classifier and helps the species classification task. Since the used dataset is too small and unbalanced for a successful supervised training, a data augmentation strategy was used. The training set in both experiments was augmented using images from the ImageNet dataset. Using images from the same species and testing the system only in our

Table 1. Mammals species sets

Set 1	Set 2
Carnivora	Didelphimorphia
Artiodactyla	Cingulata
Perissodactyla	Pilosa
	Rodentia
	Lagomorpha

Table 2. Training set before and after data augmentation

Exp	Testing set	Training set class 1	Training set class 2	Training set using ImageNet class 1	Training set using ImageNet class 2
1	100	67	1305	2193	2193
2	420	577	1600	1600	1600

dataset allowed us to successfully train supervised models. Table 2 summarizes the training and testing sets before and after using the Imagenet images.

In both experiments all deep architectures were used. Table 3 shows the six very deep ConvNets used in this work. They are the state of the art in object recognition. Since data augmentation puts a lot of the ImageNet images in the training set, a fine-tuning procedure did not have a significant effect on the system's performance, hence, just the multi-class logistic classifier at the end of the network was trained and the ConvNet was used as black box feature extractor. This work uses multiple very deep ConvNets in order to probe how the Depth in ConvNets impacts on the camera-trapping classification problem in low quality images.

4.3 Metrics

A confusion matrix is used as performance evaluation in both experiments. Efficiency, recall, specificity, and precision are extracted from the confusion matrix. In alleviation experiments only efficiency is used.

Table 3. Architectures used in the experiments

Label	Architecture	# layers
A	AlexNet	8
B	VGG Net	16
C	GoogLenet	22
D	ResNet-50	50
E	ResNet-101	101
F	ResNet-152	152

4.4 Implementation Details

All the dataset images were resized to fit in the ConvNet topologies input: AlexNet (227×227), VGGNet (224×224), GoogLenet (224×224), and ResNets (224×224). To use ConvNets as feature extractors the last full connected layer was modified to deal with 2 classes instead of 1000 Imagenet challenge classes.

All used architectures were pre-trained with the ImageNet dataset [13]. The implementation was done in the deep learning framework Caffe [14], as well as all pre-trained models that were found in the Caffe model Zoo and where performed using a Graphics processing unit Nvidia gtx 850M.

5 Results

In Table 4 the results are shown. Each experiment (1 or 2) is paired with a deep architecture (represented with a label listed in Table 3).

Table 4. Results of experiments and performance metrics

Experiment	Efficiency [%]	Recall [%]	Precision [%]
1A	77.50	80.23	69.23
1B	82.50	83.66	78.26
1C	85.00	86.63	75.00
1D	95.00	99.99	78.26
1E	**97.50**	100	90.00
1F	95.00	100	85.73
2A	74.28	70.81	75.51
2B	73.57	64.59	78.94
2C	73.09	67.46	75.40
2D	85.23	85.64	84.03
2E	**90.23**	95.69	85.83
2F	84.76	90.90	80.50

The results of Experiment 1 show a high efficiency using ResNet-101. This shows how successfully the deep architecture has learned the concept of "bird". Notice that the deeper architecture does not have the highest performance, however, the tendency depth versus performance is crescent till 101 layers. ResNet-101 has the best efficiency in Experiment 2, too. Although the performance is not as high as in Experiment 1, it is promising and even acceptable in the total automatic classification mode. In contrast to Experiment 1, the results from GoogleNet and VGGnet were worse than the results from AlexNet. These performance results can be enhanced using the alleviation mode.

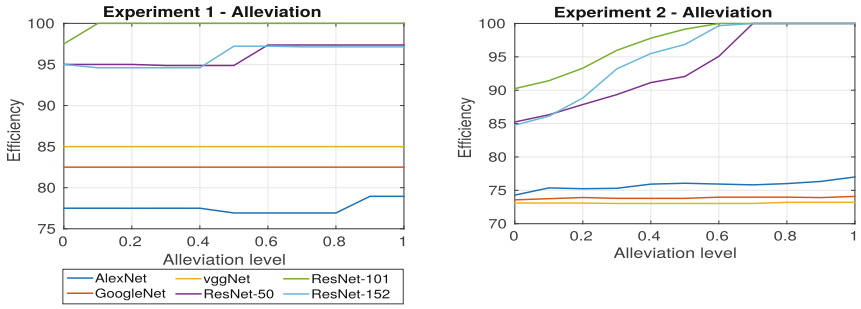

Fig. 4. (a) Alleviation in Experiment 1. (b) Alleviation in Experiment 2.

Figure 4 shows the results of the alleviation experiments. Alleviation level equal to 0 means total automatic system and increasing threshold means increasing the distance between probabilities of memberships to be an acceptable result. The alleviation did no show significance improvement in Experiment 1. Only one architecture (ResNet-101) reaches 100%, others did not improve or just a bit. An important observation is that only residual architectures and AlexNet improved using the alleviation. The residual architecture improved in a lower value of the alleviation. Experiment 2 showed a lot of improvement using the alleviation strategy. Residual architectures reached 100% of efficiency. Notice that the two deep residual nets reached in 0.6 and the less deeper near to this alleviation level 0.7. Similar to Experiment 1 the three less deep architectures did not show a lot of improvement using the alleviation and GoogleNet and VGGnet keep constant performance.

The alleviation mode represents how much the expert trusts in the classifier. When the expert expects high confidence in the classifier's decision and the model has high learning capacity (e.g. deeper networks) a perfect performance is reached.

6 Conclusions

In this paper a two step classification strategy using deep convolutional neural networks for low quality camera-trap images was proposed. Instead of directly classifying species (which previously showed low performance in this type of images) the images are classified in two steps: first birds are separated from mammals, then mammals are classified in two sets proposed by biologists. As proved in this work cluster of mammal species reduce the complexity of the classification task and still give a lot of information to the biologists. In the first classification phase the best performance was 97.5% using residual networks with 101 layers. In the second step the best performance was 90.23% also using residual networks with 101 layers. An alleviation mode is proposed to hand over

difficult classification decisions to human experts. In both classification task the alleviation mode allows the model to reach 100% of efficiency.

In the future the system will be tested using all images of the same burst (temporal information), hence the system has more opportunity to predict the correct class. Also when more data is captured a split version dataset according to Fig. 2 will be evaluated.

References

1. O'Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera Traps in Animal Ecology: Methods and Analyses. Springer Science & Business Media, New York (2010)
2. Fegraus, E.H., Lin, K., Ahumada, J.A., Baru, C., Chandra, S., Youn, C.: Data acquisition and management software for camera trap data: a case study from the team network. Ecol. Inform. **6**, 345–353 (2011)
3. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. **2013**, 1–10 (2013)
4. Chen, G., Han, T.X., He, Z., Kays, R., Forrester, T.: Deep convolutional neural network based species recognition for wild animal monitoring. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 858–862. IEEE (2014)
5. Gomez, A., Salazar, A., Vargas, F.: Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. arXiv preprint arXiv:1603.06169 (2016)
6. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 148–156 (1994)
7. Beijbom, O., Edmunds, P.J., Roelfsema, C., Smith, J., Kline, D.I., Neal, B.P., Dunlap, M.J., Moriarty, V., Fan, T.Y., Tan, C.J., et al.: Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. PloS one **10**, e0130312 (2015)
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**, 211–252 (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. arXiv preprint arXiv:1408.5093 (2014)