Towards Automatic Identification of Elephants in the Wild

Matthias Körschens, Björn Barz, Joachim Denzler
Computer Vision Group,
Friedrich Schiller University Jena
{matthias.koerschens,bjoern.barz,joachim.denzler}@uni-jena.de

Abstract

Identifying animals from a large group of possible individuals is very important for biodiversity monitoring and especially for collecting data on a small number of particularly interesting individuals, as these have to be identified first before this can be done. Identifying them can be a very time-consuming task. This is especially true, if the animals look very similar and have only a small number of distinctive features, like elephants do. In most cases the animals stay at one place only for a short period of time during which the animal needs to be identified for knowing whether it is important to collect new data on it. For this reason, a system supporting the researchers in identifying elephants to speed up this process would be of great benefit.

In this paper, we present such a system for identifying elephants in the face of a large number of individuals with only few training images per individual. For that purpose, we combine object part localization, off-the-shelf CNN features, and support vector machine classification to provide field researches with proposals of possible individuals given new images of an elephant.

The performance of our system is demonstrated on a dataset comprising a total of 2078 images of 276 individual elephants, where we achieve 56% top-1 test accuracy and 80% top-10 accuracy.

To deal with occlusion, varying viewpoints, and different poses present in the dataset, we furthermore enable the analysts to provide the system with multiple images of the same elephant to be identified and aggregate confidence values generated by the classifier. With that, our system achieves a top-1 accuracy of 74% and a top-10 accuracy of 88% on the held-out test dataset.

1 Introduction

In biological research projects, there often is a focus on certain individuals from a larger group. These individuals might be particularly interesting, be it because they are the matriarch of a huge family, or simply because they have a certain special behavior, not commonly found among others of its kind, and thus it is important to collect new data from them. As many of the individuals might look very similar, it can be very hard to identify them on the spot.

Figure 1: An example classification containing the automatically detected head bounding box and the predicted name of the elephant.

Sometimes researchers are staying with such animals for a long time, get familiar with them and thus also become able to identify them without any notes taken before. But as this is most often not the case, and most researchers often only stay there for shorter periods of time, they are not able to get that familiar with the animals and thus have to go through all the archived features concerning each individual for identification. This is not only a very time-consuming but also exhausting task. Additionally, the need for manual identification can lead to stagnation of research, because the animals of interest may move to another place before they can be identified, so that valuable data that could have been collected about them is lost.

This can slow down the research progress significantly. Thus, it would be very advantageous to have a supportive system, which can help researchers to identify the individuals quickly. In this work we propose to achieve this by combining object localization for finding elephant heads, off-the-shelf CNN features as descriptors, and support vector machines (SVMs) for multi-class classification. An example identification of an elephant from our dataset is shown in Figure 1.

The remainder of this paper is organized as follows: We briefly review related work in section 2 and describe our proposed system for identification of individual animals in section 4, after introducing the elephant dataset used for our work in section 3. Experimental results are presented in section 5 and, after a short problem analysis in section 6, we will conclude this work in section 7.

2 Related Work

In the context of human beings, face identification is a very actively studied field, where breakthroughs have recently been achieved using deep learning with systems trained end-to-end, *e.g.*, FaceNet [Schroff *et al.*, 2015], VGG-Face [Parkhi *et al.*, 2015], or DeepFace [Taigman *et al.*, 2014]. However, such approaches usually require large amounts of annotated training images per class, which are often not available in wildlife monitoring scenarios.

Loos *et al.* hence used traditional face identification methods such as Eigenfaces and SVMs for identifying 25 individual chimpanzees [Loos *et al.*, 2011] and later extended this by automatic face detection [Loos and Ernst, 2013].

Brust *et al.* have recently taken this approach to the deep learning age for gorilla identification using pre-trained convolutional neural networks (CNNs) for face detection and feature extraction [Brust *et al.*, 2017]. Their approach is, in principle, very similar to ours. However, we do not only demonstrate that it is also suitable for identifying other species such as elephants, but also show that the performance can be improved further by using earlier layers than the last layer of a CNN for feature extraction and additional pooling. Moreover, we found that simple data augmentation such as flipping can be useful for training the SVM classifier and show how to aggregate predictions obtained for multiple images of the same unknown individual to deal with occlusion and variations in pose and perspective.

In contrast to the formerly mentioned works, our dataset also poses new challenges: a large number of classes, a very small and imbalanced number of images per class, and a long time period during which images have been taken (17 years).

3 Elephant Dataset

We use a dataset provided by biologists from Cornell University, who run research on elephants in the Kongo in a project called *The Elephant Listening Project*[1], especially in the region of the Dzanga-Sangha special reserve [Turkalo *et al.*, 2017]. This research has been going on since 1999, mostly focusing on one clearing many different elephants come to every year.

Over the years, about 4000 different elephants were sighted there and documented. Many elephants have been given names and, of course, photos and videos of them were taken. Distinctive features, which can be used to identify the elephants, were documented. Since those features can be very subtle or even change over time, identifying the individual elephants can be very hard. A reliable and fast identification system would hence be of great benefit for their research.

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

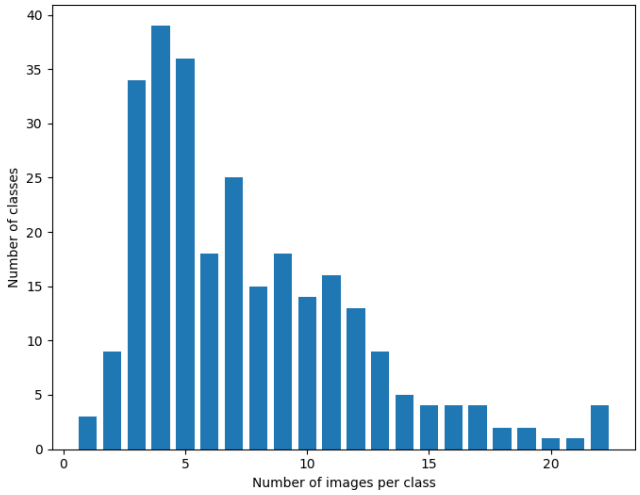

Figure 2: Number of classes with a certain amount of images. The minimum number of images in one class is 1, and the maximum 22.

The dataset consists of 2078 images of 276 different elephants. This results in about 8 images per class on average, but as we can see in Figure 2, the images are not very evenly distributed across the classes. The maximum number of images in a class is 22 and the minimum number is one. We can also see that a big part of the classes only has three to five images, which in turn results in only two to four images for training for these classes.

For our experiments we divided the dataset into a stratified training split of 75% of the images, and a corresponding test set with 25%. This results in 1573 images for training and 505 for testing.

4 Methods

The processing pipeline of our proposed approach is illustrated in Figure 3. Initially, the user inputs one or multiple images. We then automatically locate the elephants' heads in these images using a YOLO network [Redmon *et al.*, 2016] that has been pre-trained on another dataset built from Flickr images of elephants. This dataset is completely disjoint from the one used for the identification experiments and consists of 1285 training and 227 testing images.

The bounding boxes are drawn around the head instead of the entire body, since preliminary results indicated that the head contains more valuable features for identification.

The predicted bounding boxes are then shown to the user, who then can correct the bounding boxes by drawing a new one, or simply selecting one of multiple proposed ones, as multiple elephants might be contained in the image.

These selected bounding boxes are then being cut out and fed into a modified ResNet50 network [He *et al.*, 2016] for feature extraction. The base network used is the Keras implementation of ResNet50, trained on ImageNet, as the number of images in our dataset is too small to fully train a deep network. It also was modified to extract features not from the last layer before the classification layer, but from earlier activation layers, which are then followed additionally by a new pooling layer to increase translation invariance.

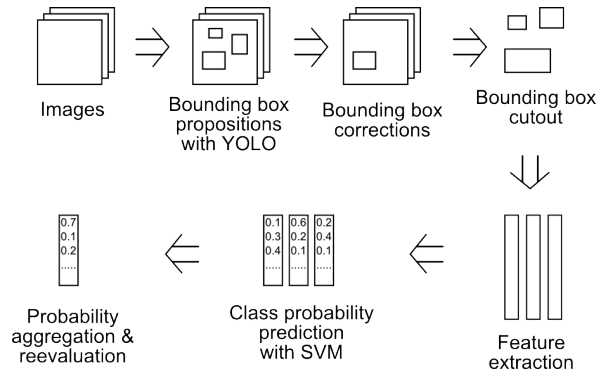

Figure 3: The full pipeline of our system from image input to the classification result.

Depending on the layer used for feature extraction and the degree of pooling, the number of features resulting from this may be extremely high, leading to long processing times and possibly memory problems. To deal with this, we apply principal components analysis (PCA [Pearson, 1901]) to reduce the number of features to twice the number of training images, *i.e.*, approximately 3,000 in the case of our 75% training split.

The extracted features are then classified using a support vector machine (SVM [Cortes and Vapnik, 1995]) and the classes, *i.e.*, the individual elephants, are sorted in decreasing order by their confidence values obtained from the SVM to create a ranking with the most probable elephant at the top. The ranking is then shown to the user, who can decide, which elephant is the most similar one to the image he provided. A few representative images from the training dataset are shown for each predicted class in the ranking, so that the user can easily filter out false positives.

To leverage the fact that some distinctive features of the elephant head are symmetric, *e.g.*, the tusks or the shape of the ears, we augment the training dataset for the SVM by adding features of horizontally flipped versions of all training images.

If multiple images were input, they are used for a joint classification, during which the confidence scores of the SVM for each single image are being aggregated by averaging the class-wise confidence values over all input images. This results in a single confidence score for each individual, so that we can proceed as in the case of a single input image.

5 Experimental Results

Object Localization Experiments with the YOLO network suggested that the head of the elephant is easier to detect, and experiments regarding the identification also suggested that head features are preferable for the classification, as they appear to have more features important for identifying the animals. Because of this, we will only focus on the elephant heads and head bounding boxes respectively in the following.

After training the network with 1285 images, its performance was tested on a test set of 227 images, resulting in a precision of 92.73% and a recall of 92.16%. The mean average precision achieved was 90.78%. Thus, we can reliably locate the position of the elephant heads in the images automatically and hence reduce the effort of manual bounding box annotation imposed on the user.

Top k	1	5	10	20
max_4 act. 40	0.508	0.706	0.770	0.823
max_5 act. 40	0.544	**0.726**	**0.8**	0.839
max_6 act. 40	**0.560**	0.716	0.788	**0.853**
max_4 act. 43	0.522	0.716	0.766	0.823
max_5 act. 43	0.546	0.708	0.770	0.833
max_6 act. 43	0.524	0.700	0.762	0.821
no pool act. 43	0.518	0.659	0.740	0.805

Table 1: Max pooling with one image using 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 $n \times n$.

Top k	1	5	10	20
max_4 act. 40	0.698	0.818	0.866	0.902
max_5 act. 40	0.714	0.832	0.876	0.904
max_6 act. 40	**0.742**	**0.852**	**0.878**	0.906
max_4 act. 43	0.700	0.830	0.874	**0.908**
max_5 act. 43	0.722	0.832	0.876	0.906
max_6 act. 43	0.708	0.828	0.868	0.904
no pool act. 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 $n \times n$.

Object Identification The identification was done using the combination of a modified ResNet50 as feature extractor and a support vector machine (SVM), which performs the actual classification. The features from the last layer before the classification layer proved to be not the best to extract features from, but the activation layer of the 14^{th} residual block in the ResNet50 architecture, here referred to as activation_43, provided better performance with the final classifier.

But this is only the case, if the features are extracted and directly being fed into the support vector machine. In contrast, when using an additional pooling layer, as described in section 4, features extracted from the 13^{th} residual block perform better, as can be seen in Table 1 and Figure 4. The abbreviations max_n stand for a max pooling layer with a pooling size of $n \times n$. We can see that the best results using pooling were achieved with the activation_40 layer, despite the 43rd layer being the best beforehand. The best top-1 accuracy is 56% with an average per-class accuracy of 49%. In the top 10 we even achieve up to 80% and 74% per-class respectively.

As the dataset is comparatively small compared to most other datasets and most classes have a much smaller number of images than the average of eight, many difficulties can occur. For example, the features needed for correct classification can be missing in the training or testing images, be-

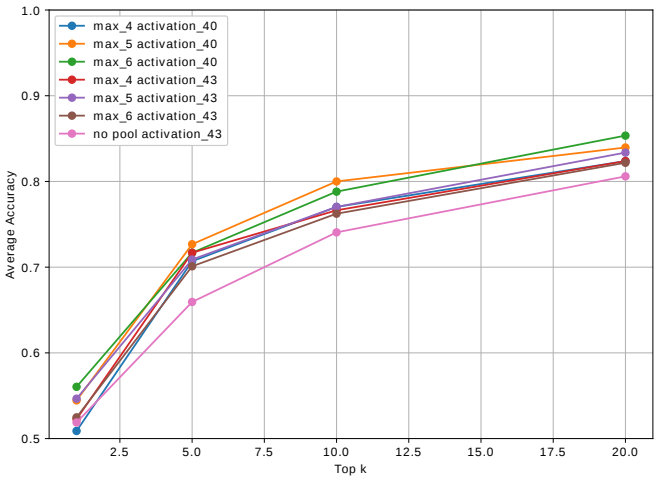

Figure 4: Comparison of different sizes of max pooling layers with single-image classification.

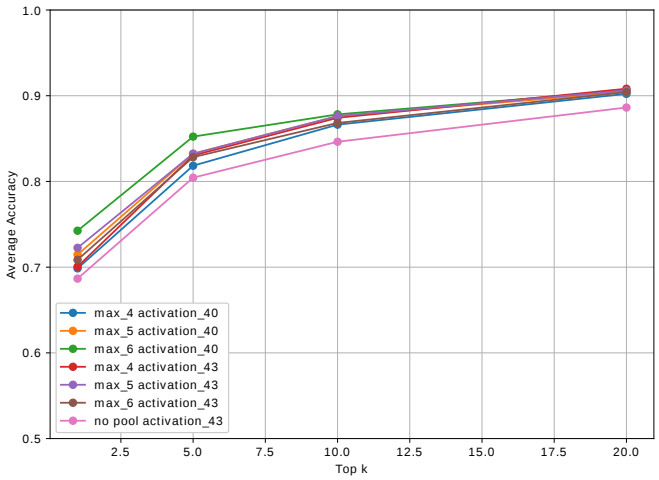

Figure 5: Comparison of different sizes of max pooling layers with two-image classification.

cause they are occluded. This can, for example, be caused by mud, or they are simply distorted due to the angle from which the picture was taken. The angle and the movement of the elephant can also have an effect on the features recorded in the picture. Lastly, it could also simply be the case that the elephant to identify was photographed from a view that has never been seen in the dataset, for example an image being input with a left side view of the elephant and the dataset containing only images of the elephant's right side.

All these cases can lead to misclassification. A solution for this problem would be to use multiple images for classification that ideally contain multiple views of the same elephant, and then combine the results of these images as described in section 4. The results of this can be seen in Table 2 and Figure 5. We can see that, using two images, an accuracy of 74% is possible and a per-class average accuracy of 59%. Among the top 10 results, we even achieve up to 88% overall and 79% per-class respectively.

From this we can conclude that it might be a good approach to use multiple images for one classification, as the combined features result in a higher accuracy. This might also be true when using multiple low-quality images, which on their own are not suited for a good classification result, but together it might be possible to still get a successful classification when combining multiple of these.

6 Problem Analysis

As we have seen during the introduction of the dataset used, the images are not distributed very evenly across the classes. This results in many misclassifications, as a lot of classes have only a small number of images to train with. But this is not the only problem we have to deal with in this dataset.

On further analysis of the images we found that some pictures are zoomed in too much, which causes important parts of the elephant's head, like the ears or the tusks, to be outside of the image and thus renders them unusable for classification. In many images the elephants are also very muddy. This often can result in the elephant being mostly monochrome in the image, which is why we cannot see some important contours, *e.g.*, the shape of the ears. Additionally, the images in the dataset have been taken in the years 2000-2016, so a further problem can be that the appearance of the animals and some of their distinctive features changed during that time. This can for example be the case, if the elephant loses a tusk or a hole in the ears becomes a rip.

7 Summary

In this project we successfully implemented a system able to assist biologists to identify elephants they encounter in the field. This system consists of a bounding box detector, implemented with YOLO, and a classifier using a combination of a modified ResNet50 as feature extractor, a PCA and an SVM as classifier. These components are connected through a pipeline and can be used via a web interface.

We can achieve about 56% top-1 and 80% top-10 accuracy, if we use one image for classification. If we use two images, we can even achieve 74% and 88% accuracy respectively. With these results the system will definitely be able to help the biologists in identifying elephants and allow them to focus more on collecting data than on identification.

There are multiple things that still can be done in the project. For example, we noticed that the results are sometimes very dependent on the bounding box drawn. To counter this, an ensemble approach using a multitude of random crops of the original bounding box as input for the classification could be used. Here, a majority vote or also an average of the confidence scores for each crop could be used to receive the actual classification.

Acknowledgments

We would like to express our thanks to Andrea Turkalo, who took the images of the elephants used in this work, and the Elephant Listening Project at the Cornell Lab of Ornithology for preparation of the material and metadata. We would also like to thank Peter Wrege and Daniela Hedwig for their valuable feedback and discussions.

References

[Brust *et al.*, 2017] Clemens-Alexander Brust, Tilo Burghardt, Milou Groenenberg, Christoph Käding, Hjalmar Kühl, Marie L. Manguette, and Joachim Denzler. Towards automated visual monitoring of individual gorillas in the wild. In *ICCV Workshop on Visual Wildlife Monitoring (ICCV-WS)*, pages 2820–2830, 2017.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

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

[Loos *et al.*, 2011] Alexander Loos, Martin Pfitzer, and Laura Aporius. Identification of great apes using face recognition. In *19th European Signal Processing Conference*, pages 922–926. IEEE, 2011.

[Parkhi *et al.*, 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015.

[Pearson, 1901] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali 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.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James 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.

[Taigman *et al.*, 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior 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.

[Turkalo *et al.*, 2017] Andrea K Turkalo, Peter H Wrege, and George Wittemyer. Slow intrinsic growth rate in forest elephants indicates recovery from poaching will require decades. *Journal of Applied Ecology*, 54(1):153–159, 2017.