A Dataset and Application for Facial Recognition of Individual Gorillas in Zoo Environments

Otto Brookes
University of Bristol, UK
Dept. of Computer Science
dl18206@bristol.ac.uk

Tilo Burghardt
University of Bristol, UK
Dept. of Computer Science
tilo@cs.bris.ac.uk

ABSTRACT

We put forward a video dataset with 5k+ facial bounding box annotations across a troop of 7 western lowland gorillas (*Gorilla gorilla gorilla*) at Bristol Zoo Gardens. Training on this dataset, we implement and evaluate a standard deep learning pipeline on the task of facially recognising individual gorillas in a zoo environment. We show that a basic YOLOv3-powered application is able to perform identifications at 92% mAP when utilising single frames only. Tracking-by-detection-association and identity voting across short tracklets yields an improved robust performance at 97% mAP. To facilitate easy utilisation for enriching the research capabilities of zoo environments, we publish the code, video dataset, weights, and ground-truth annotations at data.bris.ac.uk.

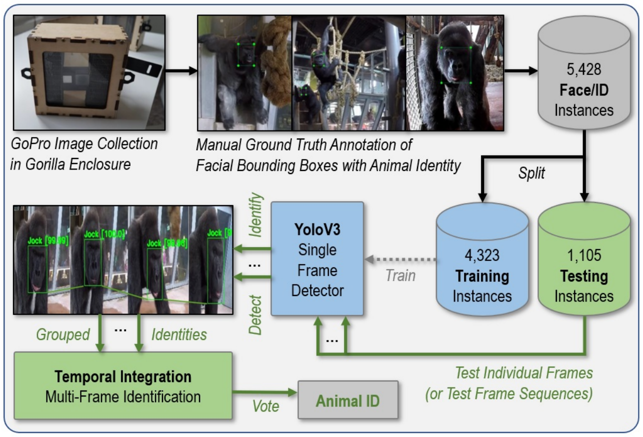

GoPro Image Collection in Gorilla Enclosure
Manual Ground Truth Annotation of Facial Bounding Boxes with Animal Identity
5,428 Face/ID Instances
Split
4,323 Training Instances
1,105 Testing Instances
Train
YoloV3 Single Frame Detector
Identify
Detect
Grouped
Identities
Temporal Integration Multi-Frame Identification
Vote
Animal ID
Test Individual Frames (or Test Frame Sequences)

Figure 1: Data Collection, Training & System Overview. (a) Ruggedly enclosed GoPro cameras were placed in a zoo enclosure housing a group of 7 lowland gorillas at Bristol Zoo Gardens. (b) 5k+ frames from the gathered video data were then manually annotated with a bounding box and identity information. (c) This dataset was used to train and evaluate a YOLOv3 detection and classification framework, which was tested as (d) a single frame recognition system, and a multi-frame system yielding (e) location and identity data of the gorilla faces in view.

1 INTRODUCTION

Motivation. One important area of focus for zoos and sanctuaries is animal welfare [4] and associated research questions: e.g. does captivity prohibit animals from functioning in a beneficial capacity [2, 14]; and how does captivity affect the ability of animals to be potentially reintroduced into the wild [8]? Answering such questions via prolonged monitoring is particularly relevant for Great Apes where many gorilla species are critically endangered [22]. Manual monitoring by specialists [7], however, is labour intensive.

Contribution. This paper provides a new annotated dataset for Great Ape facial ID and investigates how far YOLOv3 [17] can be used to simultaneously detect and identify individual zoo gorillas based on facial characteristics (see Fig. 1). Our contributions are: (1) Collection and annotation of 5,428 samples of 7 western lowland gorillas (see Fig. 2); (2) Training and evaluation of the YOLOv3 framework for single frame gorilla face localisation and classification, and (3) Implementation of an offline multi-frame video application that delivers robust IDs in a zoo environment.

2 BACKGROUND & RELATED WORK

Human Facial Biometrics. Facial recognition technology for humans has long been a key field within computer vision [20, 21]. Deep convolutional neural networks (CNNs) [11] exemplified in frameworks such as DeepFace [19] form the base of most modern facial biometric frameworks [15].

Great Ape Face Recognition. Great Apes show facial characteristics that are individually distinct (see Fig.2) and not dissimilar to those of humans owing to our close evolutionary lineage [13]. Thus, methodologies in animal biometrics [10] follow approaches for human face recognition closely. Loos *et al.* [12] developed one of the first chimpanzee facial recognition systems based on traditional machine learning techniques. He also annotated key datasets [12], where further datasets have since been presented by Brust *et al.* [1] and Schofield *et al.* [18]. Building on Loos's work and data, Freytag *et al.* [5] trained a deep learning object detector, YOLOv2 [16], to localise the faces of chimpanzees. They utilised a second deep CNN for feature extraction (AlexNet [9] and VGGFaces [15]), and a linear support vector machine (SVM) [3] for identification. Later, Brust *et al.* [1] extended their work utilising a much larger and diverse dataset. Most recently, Schofield *et al.* [18] presented a pipeline for identification of 23 chimpanzees across a video archive spanning 14 years. Similar to Brust *et al.* [1], a single-shot object detector, SSD [18], is trained to perform localisation, and a secondary CNN model is trained to perform individual classification. They group video detections into tracklets across which identities are computed from, showing an improvement over single frame operation. In contrast to all discussed systems, in this paper we employ YOLOv3 [17] to perform *one-step* simultaneous facial detection and individual identification on gorillas, leading to a simpler yet equally effective

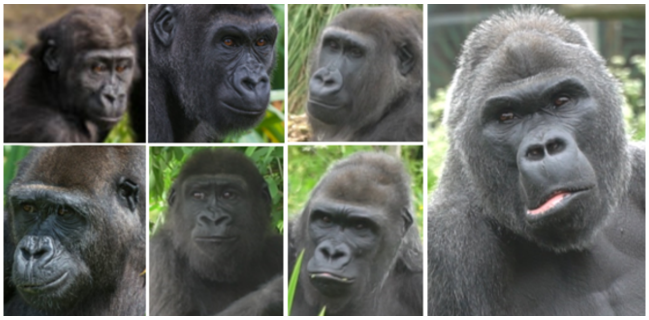


Figure 2: Lowland Gorilla Troop. The figure depicts row-by-row left-to-right: Ayana, Kukuena, Kala, Touni, Afia, Kera. The large image on the far right is of Jock. The troop consists of 1 male and 6 females aged between 6 months and 37 years.

pipeline. Details of the existing great ape facial recognition systems and datasets can be found in the following literature; Loos *et al.* 2011 [13], Loos *et al.* 2013 [12], Freytag *et al.* 2016 [5], Brust *et al.* 2017 [1] and Schofield *et al.* 2019 [18]

3 DATASET

BristolGorillas2020. The BristolGorillas2020 dataset comprises 628 video segments and 5,428 annotated facial images (sampled from the corresponding video segments). The train-test splits for the images were generated using stratified sampling (see Table 1). The test set for the single-frame and multi-frame detector are the same; each tracklet comprises detections made on the *n* frames preceding the ground-truth frames included the test set. Data used in training is not evaluated during testing of the video application. The dataset GoPro (versions 5 & 7) and Crosstour Action cameras (see Fig. 1) were fitted in the enclosure near enrichment devices [6] to obtain close-up facial footage of the gorillas (see Fig. 3). Data collection took place twice per week (from 11am to 1pm) over 6 weeks to record RGB video at 1280×720 pixels and 30fps. A selection of frames containing front facial images was manually labelled (see Fig. 3, top) with the help of experts from the Primate division at Bristol Zoo to ensure the identities of individual gorillas were labelled correctly.

Table 1: Complete Dataset. The table shows the total number of collected facial image patches, including the number of training and testing patches for each individual gorilla.

Individual Gorillas	Training	Testing	Total Images
Afia	614	157	771
Ayana	489	126	615
Jock	387	101	488
Kala	578	148	726
Kera	776	196	972
Kukuena	747	190	937
Touni	732	187	919
Total	**4,323**	**1,105**	**5,428**

4 IMPLEMENTATION

Single-Frame Identification. To train YOLOv3 we employed the open-source and freely available Darknet implementation pre-trained on ImageNet1000 [17]. This network was then fine-tuned using stochastic gradient descent with momentum and a batch size of 32 at an input resolution of 416×416 RGB pixels. Fine-tuning was performed with batch normalisation, data augmentation (see Fig. 3, bottom), and learning rate decay (an initial learning rate of 0.001 reduced by a factor of 10 at 80% and 90% of the total training iterations). We trained against a one-hot encoded class vector, where each vector element represented a gorilla identity. The resulting YOLOv3 network forms the backbone of the facial recognition system by performing one-step multi-instance localisation and identification of gorilla faces.

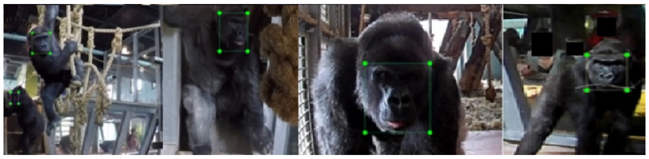


Figure 3: Gorilla Face Annotation. A collection of images annotated with front facial bounding boxes.

Multi-Frame Identification. In Phase 2 the network is applied to individual frames of a sequence (i.e. tracklets across video) where X_t denotes the frame at time step t. All detections in X_t and X_{t+1} are then input into a simple algorithm that uniquely associates cross-frame detections if they show the highest pairwise intersection-over-union (IoU) and this IoU is also greater than a threshold $\theta = 0.5$. The resulting association chains represent tracklets (see Fig. 4). Their length ranges from a single frame to 10 frames. For each tracklet we evaluate identity classification via two methods: (1) highest single class probability denoted as *maximum*, or the highest average class probability denoted as *average.* For each detection we use the IoU assignment to compare the ground truth against the identity assigned to the tracklet containing the detection (where non-detections contribute to false negatives).

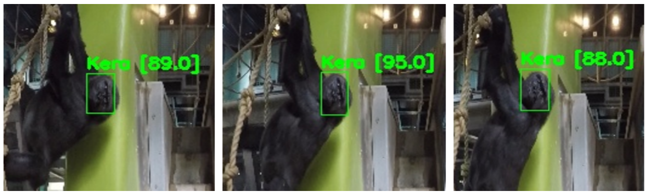
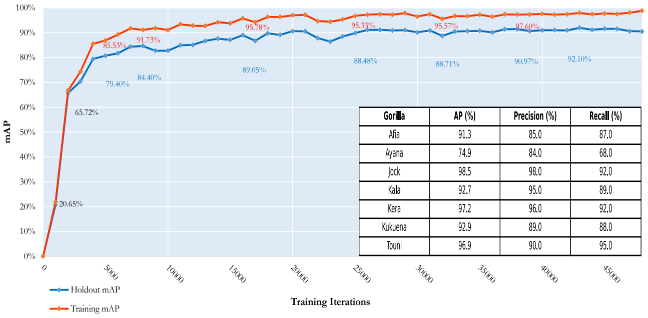


Figure 4: Multi-Frame Recognition. The figure exemplifies a tracklet and shows confidence scores for the dominant class 'Kera' on a sequence where she swings across from a nest.

5 RESULTS

Single Frame Detector. Figure 5 reports single frame classification performance of YOLOv3 as precision, recall, and mean average precision (mAP) for the test set as well as mAP curves across the training process over 44k steps for both training (red) and test (blue) sets. We noted that the average IoU of this network was 77.92% (see Fig. 6), whilst the average IoU scores for Afia and Ayana are only 71% and 67%, i.e. 6.92% and 10.92% below the network average, respectively. These are younger members of the troop; they are physically smaller in size and are often in the background rather than directly in front of the camera. This clearly challenged the network with regard to precise localisation and identification. Despite

YOLOv3's improved detection capabilities [17], ID classification suffers if image patches are too small. For instance, Ayana's AP score at 74.9% is particularly low which may in part to be due to the fact that 64.34% of her ground-truth annotations are smaller than 32 × 32 pixels.



Gorilla	AP (%)	Precision (%)	Recall (%)
Afia	91.3	85.0	87.0
Ayana	74.9	84.0	68.0
Jock	98.5	98.0	92.0
Kala	92.7	95.0	89.0
Kera	97.2	96.0	92.0
Kukuena	92.9	89.0	88.0
Touni	96.9	90.0	95.0

Figure 5: Single Frame YOLOv3 Identification Performance. The graph shows the mAP scores on training (red) and testing (blue) sets against training iterations. The table depicts testing results for each of the individuals.

Multi-Frame Detector. Table 2 reports multi-frame classification performance via precision, recall, and AP for the test set, where network weights are used which achieved the highest mAP score in single frame detection. The results reported utilise voting across a maximum tracklet size of 5, a stride of 1 and an IoU association threshold of 0.5. The multi-frame detector with maximum voting achieves the highest mAP, however, there is only a marginal difference between the maximum and average voting algorithms with less than 0.5% difference between all three of the reported evaluation metrics. Both multi-frame detection approaches outperform the single frame detector across all metrics. The mAP improvements achieved by the average and maximum voting algorithms when compared with the single-frame detector are 5.2% and 5.4%, respectively.

Cross Validation. We perform stratified 5-fold cross-validation on both single-frame and multi-frame identification systems. We train each fold for 24,000 iterations owing to time and computational restrictions. The three identification systems, single-frame and multi-frame identification with average and maximum voting schemes, achieve 89.91%, 95.94% and 96.65% mAP, respectively.

Table 2: Multi-Frame Detector Performance. Performance is shown for both average and maximum voting schemes. The performance of the single-frame detector is included for comparison.

Detection	mAP (%)	Precision (%)	Recall (%)
Single	92.1 (± 8.0)	91.0 (± 5.5)	87.3 (± 9.9)
Average	97.3 (± 2.5)	95.1 (± 4.7)	91.1 (± 6.5)
Maximum	97.5 (± 2.2)	95.4 (± 2.7)	91.2 (± 7.9)

6 ABLATION STUDIES

Ablation Experiments. We now investigate the effect of tracklet size, stride and association threshold on mAP performance. Produced based on evaluation on the test set the results are shown in Table 3.

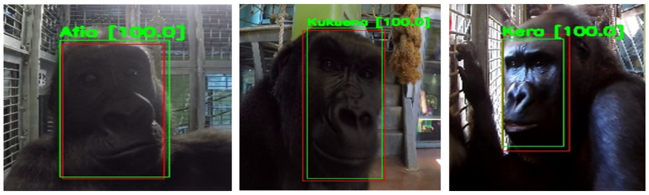


Figure 6: Localisation. The figure shows examples of ground-truth bounding boxes (red) and predicted bounding boxes (green) for Afia (left), Kukuena (middle) and Kera (right).

All three parameters prove stable with respect to changes. This indicates that the multi-frame identification improvement observed for the task at hand essentially rests with sampling multiple frames irrespective of the details of sampling.

Table 3: Ablation Experiments for Multi-Frame Detector. The table presents results produced using the multi-frame detection approach with maximum and average voting schemes varying maximum tracklet length, temporal tracklet stride, and IoU association threshold for generating tracklets, respectively.

	mAP (%)		
Tracklet Length	3 frames	5 frames	10 frames
Average	97.2 (± 2.4)	97.3 (± 2.5)	97.4 (± 2.5)
Maximum	97.4 (± 2.9)	97.5 (± 2.2)	97.5 (± 2.9)
Tracklet Stride	1 frame	3 frames	5 frames
Average	97.3 (± 2.5)	96.4 (± 3.4)	96.5 (± 3.3)
Maximum	97.5 (± 2.2)	97.4 (± 3.4)	97.5 (± 3.3)
Association Threshold	IoU>0.25	IoU>0.5	IoU>0.75
Average	97.1 (± 2.7)	97.3 (± 2.5)	96.7 (± 2.6)
Maximum	97.5 (± 2.9)	97.5 (± 2.2)	97.5 (± 2.7)

7 QUALITATIVE DISCUSSION

Detection Errors. Most of the observed false-positives are attributable to ground-truths with a small bounding box (less than 32 x 32-pixel resolutions). The majority relate to either Afia or Ayana, the youngest members of the troop, who rarely appear in the forefront of footage and have the largest proportion of small ground truth bounding boxes. This suggests that YOLOv3 is less effective at detecting small objects although there are many other factors to consider. However, most of the remaining false-positive detections appear to be caused by extreme variations in pose.

8 CONCLUSION

Summary. We presented a new dataset for the facial identification of Gorillas (*Gorilla gorilla gorilla*) which contains 5k+ annotated frames. We evaluated a multi-frame deep learning system based on the YOLOv3 framework on the task of facially recognising 7 individual western lowland gorillas in a zoo environment. We showed that the system is able to perform identifications above 92% mAP when operating on single frames and above 97% mAP when operating on tracklets. We conclude that, despite the availability of more complex systems, the proposed straight forward end-to-end application as presented operates sufficiently robustly to be used for enriching the research capabilities of zoo environments as well as their visitor experience.

Future Work. We intend to train our system on the dataset compiled by Brust et al [1] and Schofield et al [18] for comparison. Furthermore, preparatory work to install our system as a permanent feature at Bristol Zoo is underway. Footage from a camera fitted in the enclosure will be streamed to a screen in the visitor area and display the identities of individual gorillas. It is hoped that this will improve the visitor experience and help to raise revenue for the zoo to invest in further research or conservation programs.

9 ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of our collaborators for their help in facilitating the technical innovation described in this paper. Specifically, we'd like to thank the other members of the Gorilla Game Lab team, Dr Fay Clark, Dr Katy Burgess, Dr Peter Bennett, and Dr Stuart Gray for their efforts. We would also like to extend our thanks to the staff at Bristol Zoo Gardens, most notably to the keepers, and to Betty Roberts.

REFERENCES

[1] Clemens-Alexander Brust, Tilo Burghardt, Milou Groenenberg, Christoph Kading, Hjalmar S Kuhl, Marie L Manguette, and Joachim Denzler. 2017. Towards automated visual monitoring of individual gorillas in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. 2820–2830.

[2] Dalia Amor Conde, Nate Flesness, Fernando Colchero, Owen R Jones, Alexander Scheuerlein, et al. 2011. An emerging role of zoos to conserve biodiversity. *Science* 331, 6023 (2011), 1390–1391.

[3] Nello Cristianini, John Shawe-Taylor, et al. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[4] David Fraser. 2009. Assessing animal welfare: different philosophies, different scientific approaches. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association* 28, 6 (2009), 507–518.

[5] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S Kühl, and Joachim Denzler. 2016. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition*. Springer, 51–63.

[6] Stuart Gray, Fay Clark, Katy Burgess, Tom Metcalfe, Anja Kadijevic, Kirsten Cater, and Peter Bennett. 2018. Gorilla Game Lab: Exploring modularity, tangibility and playful engagement in cognitive enrichment design. In *Proceedings of the Fifth International Conference on Animal-Computer Interaction*. 1–13.

[7] Kiersten Austad Jarvis. 2007. *Effects Of a Complex Enrichment Device On Tool Use, Tool Manufacturing, Activity Budgets, And Stereotypic Behaviors In Captive Western Lowland Gorillas (Gorilla gorilla gorilla)*. Ph.D. Dissertation. University of West Florida.

[8] Devra G Kleiman, Katerina V Thompson, and Charlotte Kirk Baer. 2010. *Wild mammals in captivity: principles and techniques for zoo management*. University of Chicago Press.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[10] Hjalmar Kuehl and Tilo Burghardt. 2013. Animal biometrics: Quantifying and detecting phenotypic appearance. *Trends in ecology & evolution* 28 (03 2013). https://doi.org/10.1016/j.tree.2013.02.013

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

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

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

[14] M Elsbeth McPhee and Kathy Carlstead. 2010. The importance of maintaining natural behaviors in captive mammals. *Wild mammals in captivity: Principles and techniques for zoo management* 2 (2010), 303–313.

[15] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep face recognition.. In *bmvc*, Vol. 1. 6.

[16] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[17] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[18] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. 2019. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances* 5, 9 (2019), eaaw0736.

[19] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[20] AS Tolba, AH El-Baz, and AA El-Harby. 2006. Face recognition: A literature review. *International Journal of Signal Processing* 2, 2 (2006), 88–103.

[21] Matthew A Turk and Alex P Pentland. 1991. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 586–591.

[22] Peter D Walsh, Kate A Abernethy, Magdalena Bermejo, Rene Beyers, Pauwel De Wachter, Marc Ella Akou, Bas Huijbregts, Daniel Idiata Mambounga, Andre Kamdem Toham, Annelisa M Kilbourn, et al. 2003. Catastrophic ape decline in western equatorial Africa. *Nature* 422, 6932 (2003), 611.