Visual Recognition of Great Ape Behaviours in the Wild

Faizaan Sakib
Department of Computer Science
University of Bristol
Bristol, UK
Email: ss16161@bristol.ac.uk

Dr. Tilo Burghardt
Department of Computer Science
University of Bristol
Bristol, UK
Email: tilo@cs.bris.ac.uk

Abstract—Whilst species detection [1] and individual identification [2] of great apes have previously been attempted with success, research into automated behaviour recognition remains a human-centric task [3]. In this paper we present a first great ape-specific visual behaviour recognition system and related data annotations for detecting core ape behaviours. Among others, these include sitting, walking, and climbing. The presented basic dual-stream model with late fusion is capable of performing multi-subject multi-behaviour recognition on apes in challenging camera trap footage. More than 180,000 frames across 500 videos from the PanAfrican dataset [4] were annotated with individual IDs and behaviour labels to end-to-end train and evaluate the system. In summary, our key contributions include a proposed system capable of an accuracy of 73.52%, along with the behaviour annotated dataset, key code and network weights.

I. Introduction

The task of behaviour recognition is concerned with the automatic identification of a subject's activity within a video. Deep convolutional neural networks (CNNs) have proven to be proficient in performing this task. However, research in this area has remained largely focused on human activity [3]. Similarly, current behaviour-annotated datasets do not extend beyond humans as its subjects [5], [6].

Yet, deep learning clearly has the potential to play a key role in substantially enhancing efforts of research and conservation of great apes. The foundations of great ape "biometrics" [7] are already in place, with effective methods for ape detection [1] and identification [2] having been established. The addition of behaviour recognition represents the next step towards a system that can truly automate the monitoring of great apes.

II. Dataset and Annotations

We utilise the PanAfrican [4] dataset as an example of 'in the wild' capture of great ape behaviour. It is one of the largest of its kind in existence, made up of thousands of hours of jungle trap camera video footage filmed in central Africa. There are two species of great apes that feature in the subset of data we use: gorillas and chimpanzees. At any given frame, the total number of apes simultaneously present may range from 0 to 8. The apes across the dataset vary in species, age, gender and size. The videos span across multiple locations, at varying time of day.

In this work we extend the dataset produced in [1] which includes ape objects with manually annotated bounding boxes. The augmented annotations created as part of this work can
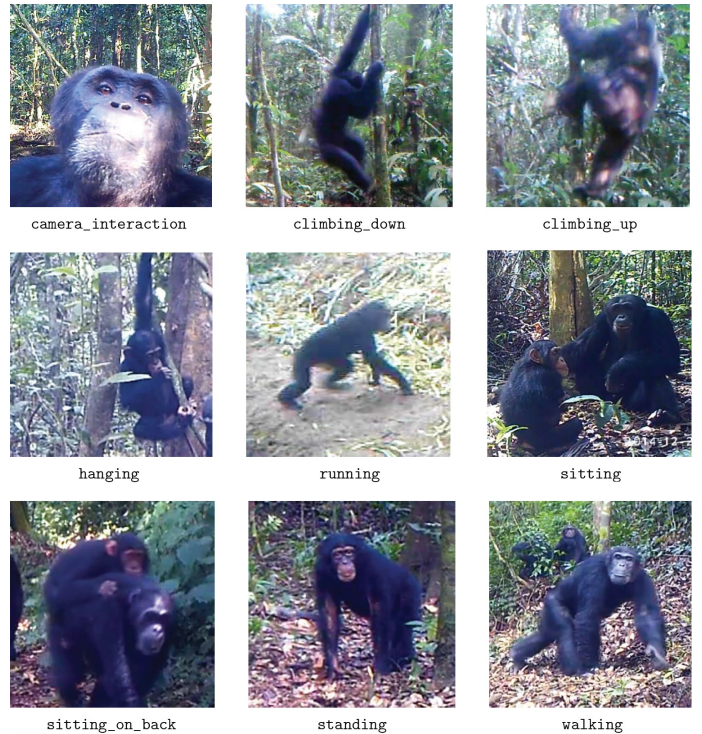

Fig. 1: **Great Ape Behaviours.** Examples of the nine core behaviours targeted in this paper. We produce a large, manually annotated training data corpus for networks to learn these behaviours and evaluate a basic dual-stream CNN architecture to demonstrate the viability of automated behaviour detection.

be found at data.bris.ac.uk under the name PanAfrican2020. In total there are $\sim 180,000$ XML annotation files, individually mapping to every frame spanning across 500 videos captured at 24 frames per seconds. We split the videos into training, validation and test sets at a ratio of 400:25:75. Scripts were written to automate the process of labelling where possible. We extended each ape object with the following metadata:

- **Behaviour:** The activity (e.g. walking) being exhibited by the great ape. This is made up of one behaviour, as simultaneous behaviours are not taken into account.
- **Identification number (ID):** A unique integer to identify each ape within a video, assigned by order of appearance starting from 0.

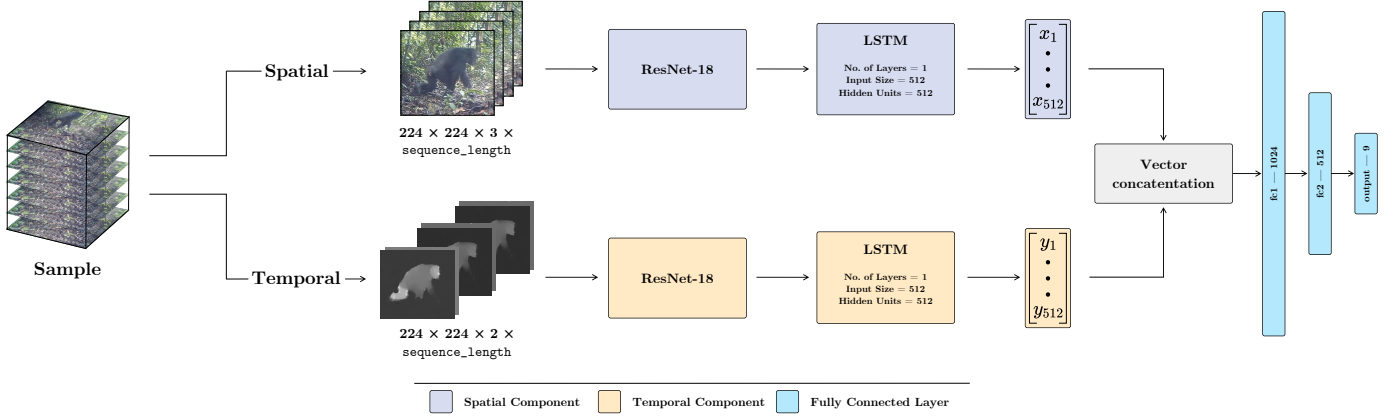A total of nine distinguishable core behaviours were

Fig. 2: **Two-Stream ResNet-18 + LSTM great ape behaviour recognition model architecture:** The spatial stream takes in a sequence of RGB frames and the temporal stream takes in a sequence of greyscale optical flow frames. An LSTM network is appended to each of the ResNet-18 CNNs. The vector output of size 512 from both LSTM networks are fused using concatenation. This is followed by a two layer fully connected network to produce the final classification scores.

identified and annotated: camera_interaction, climbing_down, climbing_up, hanging, running, sitting, sitting_on_back (a child ape mounted on an older ape's back), standing and walking as shown in Fig. 1. Meanwhile, the ID label establishes tracklets for each ape present throughout the duration of a video.

Note that there were "auxiliary" behaviours, namely eating and the act of scavenging, that were usually seen while other behaviours were already being exhibited (e.g. an ape could eat while sitting or walking). However, these were not considered as core behaviour classes since the model defined for this project aimed to only classify one behaviour at any one time.

III. METHODS

A two-stream approach, akin to the work of Simonyan et al. [8], was utilised as the backbone of the behaviour recognition model. It consists of two CNNs taking into account the spatial and temporal dependencies of a video separately (see Fig. 2). They take in a sequence of n consecutive RGB frames and n grayscale optical flow images respectively.

To fulfil the multi-subject multi-behaviour requirements of the model, we consider sampling every ape present in a given video. A *sequence length* of 20 consecutive frames is used as input to both streams. A *sampling rate* of 20 is used, effectively leaving no interval between sequences sampled. Finally, a *behaviour duration threshold* is considered, signifying the number of consecutive frames an ape has to exhibit a behaviour for it to qualify for sampling. The threshold helps filter the data to obtain quality samples showing sustained displays of behaviour where prominent characteristics are likely to be exhibited. We use a threshold of 72 frames, equalling 3 seconds.

For each valid sample, we crop the bounding boxes around the ape as per the specified ID before assembling them into sequences to be used as input to the model.

Features are then extracted from the sequence input with the use of ResNet-18 [9] CNNs. Both spatial and temporal

ResNet-18 CNNs are pretrained on ImageNet [10] and subsequently fine-tuned on our dataset to reap the benefits of transfer learning. Long short-term memory (LSTM) networks are appended to the CNN of each stream. The feature map of size 512 output from each of the CNNs are used as input for the LSTM networks.

In the case where there is no tracklet information present, detectors such as YOLO [11] can be easily trained and integrated. Additionally, the domain-specific system for ape detection proposed by Xinyu et al. [1] is a viable option.

The output from each of the LSTM networks are fused with vector concatenation amounting to a late fusion approach. Finally, this combined vector is passed through a two-layer fully connected network to obtain classification scores.

Balanced sampling was used to address class imbalance found in the dataset, by ensuring every training batch consists an equal number of samples for each class. Additionally, focal loss [12] is used as an alternative to cross-entropy loss.

The learning rate is set to 0.0001. L_2 regularisation of 0.01 is used. The focal loss is used with parameters $\alpha = 1.00$ and $\gamma = 1.00$. The batch size is set to 9 (reduced to the lowest size possible while using balanced sampling to counter the high memory requirements of the LSTM networks). Both LSTM networks use 1 layer and 512 hidden units each. The SGD optimiser is used with momentum set to 0.9.

IV. EXPERIMENTS AND BASIC RESULTS

Simple Baseline Model. Initially, a baseline model was tested with the use of ResNet-18 CNNs pretrained on ImageNet [10] in a two-stream architecture. We found that ResNet-18 proved to show good performance while retaining a relatively low level of complexity, using only ~**8.7%** of parameters when compared to VGG-16. In addition, the pretrained weights were clearly beneficial for the training process due to an increased final accuracy. The impact of pretraining on the spatial stream appeared to be much greater than on the temporal stream. Following Feichtenhofer et al. [13] we tested both late fusion and convolutional fusion to achieve
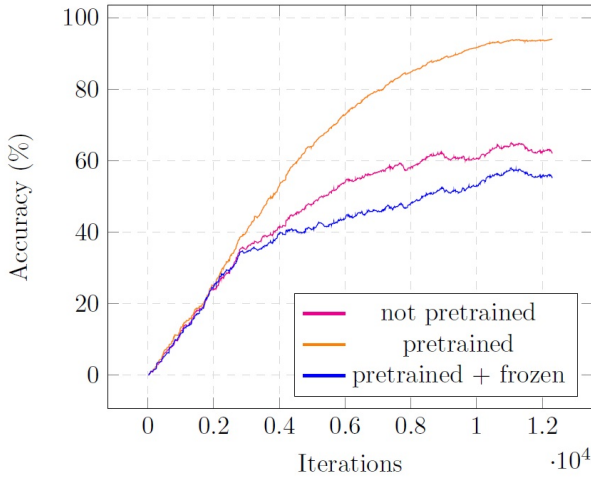
Fig. 3: **Transferability of ImageNet Weights to Spatial Stream of Recognition Model.** The graph shows accuracy during a training run with and without using pretrained weights from ImageNet to initialise the spatial stream of the network architecture. The large gains in training accuracy with the use of pretrained weights can be observed.

dual-stream convergence, where convolutional fusion comprehensively captured spatio-temporal dependencies leading to overall better performance. This baseline model featuring the design choices described above performed with a Top1 accuracy of **61.39%**, Top3 accuracy of **86.34%**.

Additional Features. Balanced sampling increased this Top1 accuracy by **7.39%**, and Top3 accuracy by **6.83%**. The parameters of the focal loss were tested, with best performance found where $\alpha = 1.00$ and $\gamma = 1.00$. LSTM networks were introduced to the model as an alternative method in capturing multi-frame information. The use of 2 LSTMs using 1 layer each without dropout performed best placing the units after the ResNet-18 CNN of each stream (see Fig. 2). Since the LSTMs output 1D vectors, using convolutional fusion which involves 3D convolutional layers was no longer suitable. Thus, we reverted back to using late fusion upon the concatenated LSTM output vectors. Table I shows the evaluation of the two models against the withheld test set, indicating the significant improvement in performance on the baseline model.

An attempt to use only one LSTM in the model for lower memory consumption was made by performing pre-LSTM fusion between the two streams. However, this resulted in a negative impact on performance. An alternate approach could be to perform spatial and temporal feature extraction using an I3D model [14] before training a single LSTM with its output.

Single Train-Test Split Results		
Model	Top1 Accuracy	Top3 Accuracy
Baseline	61.39%	86.34%
Optimised	**74.82%**	**95.49%**

TABLE I: **Single Train-Test Split Results.** The introduction of balanced sampling, focal loss and LSTM into the baseline architecture leads to a significant performance increase.

We experimented with a sequence length of 5, 10, 20, 30 frames. We were limited to 30 frames due to memory constraints. The model's performance with a sequence length of 20 frames was found to be the best, as a subsequent increase to 30 frames showed a significant drop in performance.

The optimised model with additional features performed with a Top1 accuracy of **74.82%**, Top3 accuracy of **95.49%**.

We also report preliminary detection results with YOLOv3 [11], performing with a mAP of **81.27%** on the test set. Results of the ape detector proposed by Xinyu et al. on the same dataset can be found in the original paper [1].

V. CROSS-VALIDATED RESULTS

We evaluated the model described in Section III using 4-fold cross validation. Each fold consisted of 375 training videos and 125 test videos. Table II displays the resulting performance measures. It can be seen that the performance seen on a single train-test split can be widely retained, suggesting the model generalises reasonably within the domain tested.

Final Cross Validated Results	
Top1 Accuracy	Top3 Accuracy
73.52%	**94.07%**

TABLE II: **Final 4-fold Cross-validated Results.** The performance of the single train-test split can be widely retained indicating good generalisability of the model to unseen data.

VI. QUALITATIVE ANALYSIS

When visually observing the model's predictions, we identified a number of trends where the model performs poorly. One of these was identified in the form of transitional behaviours. They are ambiguous periods where an ape is in the process of ending its current behaviour and beginning to display the next. Fig. 4 (bottom row) illustrates an example skim of such a transition.

The model is often unable to predict the behaviour of `running`, primarily mistaking it as `walking`. This suggests that the model may be ineffective in picking up the difference in speed between the two behaviours. There are also occasions where an ape stands stationary, while using their arms to dig or pick objects up. This movement, while standing, closely resembles the initial motions of walking. As such, the model predicts that the ape is starting to walk, when it is not.

VII. CONCLUSION

In this paper we introduced the first bounding-box level behaviour meta-dataset for great apes covering 180,000 frames across 500 videos of the PanAfrican camera trap archive [4]. We evaluated a basic two-stream deep learning model as a baseline using this challenging footage taken in the wild. By exploring the use of balanced sampling, focal loss and LSTMs we were able to significantly improve results. The outcome of a 4-fold cross validation procedure shows that the final model is capable of classifying nine core behaviours among gorillas and chimpanzees, at a Top1 accuracy of **73.52%**. Future work

Fig. 4: **Examples Output of the Model.** Rows show video skim sequences from the test set. Each ape is indicated by its bounding box, with its ID and detected behaviour labelled on the top left. A red box signifies an incorrect prediction, with the true behaviour appended on the label. The bottom row shows an instance of a transitional behaviour where the model fails to recognise the start of the ape walking in time.

will focus on cross-fertilisation of behaviour detection with recent advances in great ape pose recognition [15].

We hope that the behaviour annotations and proof of concept provided in this paper can aid research of the biological and conservational sciences in order to further our understanding and ultimately the protection of these charismatic species.

REFERENCES

[1] X. Yang, M. Mirmehdi, and T. Burghardt, "Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 255–262.

[2] D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho, "Chimpanzee face recognition from videos in the wild using deep learning," *Science advances*, vol. 5, no. 9, 2019.

[3] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[4] Max Planck Institute for Evolutionary Anthropology, "Pan African Programme: The Cultured Chimpanzee," http://panafrican.eva.mpg.de/index.php.

[5] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijaya-narasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[6] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[7] H. S. Kühl and T. Burghardt, "Animal biometrics: quantifying and detecting phenotypic appearance," *Trends in ecology & evolution*, vol. 28, no. 7, pp. 432–441, 2013.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2980–2988.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[14] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[15] A. Sanakoyeu, V. Khalidov, M. S. McCarthy, A. Vedaldi, and N. Neverova, "Transferring dense pose to proximal animal classes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5233–5242.