Count, Crop and Recognise: Fine-Grained Recognition in the Wild

Max Bain[1†], Arsha Nagrani[1], Daniel Schofield[2] and Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford
[2]Institute of Cognitive & Evolutionary Anthropology, University of Oxford

Abstract

The goal of this paper is to label all the animal individuals present in every frame of a video. Unlike previous methods that have principally concentrated on labelling face tracks, we aim to label individuals even when their faces are not visible. We make the following contributions: (i) we introduce a 'Count, Crop and Recognise' (CCR) multi-stage recognition process for frame level labelling. The Count and Recognise stages involve specialised CNNs for the task, and we show that this simple staging gives a substantial boost in performance; (ii) we compare the recall using frame based labelling to both face and body track based labelling, and demonstrate the advantage of frame based with CCR for the specified goal; (iii) we introduce a new dataset for chimpanzee recognition in the wild; and (iv) we apply a high-granularity visualisation technique to further understand the learned CNN features for the recognition of chimpanzee individuals.

1. Introduction

Recognising animal individuals in video is a key step towards monitoring the movement, population, and complex social behaviours of endangered species. Traditional individual recognition pipelines rely extremely heavily on the detection and tracking of the face or body, both for humans [6, 11, 18, 27, 34, 42, 56, 61, 64] and for other species [13, 48, 52, 60]. This can be a daunting annotation task, especially for large video corpora of non-human species where custom detectors must be trained and expert knowledge is required to label individuals. Furthermore, often these detectors fail for animal footage in the wild due to the occlusion of individuals, varying lighting conditions and highly deformable bodies.

Our goal in this paper is to automatically label individuals in every frame of a video; but to go beyond face and body recognition, and explore identification using the entire frame. In doing so we analyse the important trade off between precision and recall for face, body and full-frame methods for recognition of individuals in video. We target the recognition of chimpanzes in the wild. Consider the performance of models at the three levels of face, body and frame (Figure 1). Face recognition now achieves very high accuracy [44, 51, 55] for humans due to the availability of very large datasets for training face detection [31, 49, 63, 65] and recognition [2, 7, 24, 32, 59]. The result is that the *precision* of recognising individuals will be high, but the *recall* may well be low, since, as mentioned above, face recognition will fail for many frames where the face is not visible. Using a body level model occupies a middle ground between face and frame level: it offers the possibility of recognising the individual when the head is occluded, e.g. by distinguishing marks or shapes in the case of animals, or by hair or clothes in the case of humans (albeit it is worth noting that changes in clothing can reduce this advantage – animals obligingly are unclothed). However, body detectors do not as yet have the same performance as face detectors, as animal bodies in particularly are highly deformable and can often overlap each other. This means that bodies may be missed in frames, especially if they are small. A frame level model offers the possibility of very high recall (since there are no explicit detectors that can fail, as there are for faces and bodies). In addition, such a method can implicitly use higher-level features for recognition, such as the co-occurrence and spatial relationships between animal individuals (eg. infants are often present in close proximity to the mother). However, the precision may be low because of the challenge of the large proportion of irrelevant information present in the frame (in the case of body and particularly face detection, irrelevant information is removed).

In this paper we show that the performance of frame level models can be considerably improved by automatically zooming in on the regions containing the individuals. This then enables the best of both worlds: cheap supervision at the frame level, obviating the necessity to train and employ face or body detectors, and high recall; but with the precision comparable to face and body detection. We

†Correspondence at maxbain@robots.ox.ac.uk

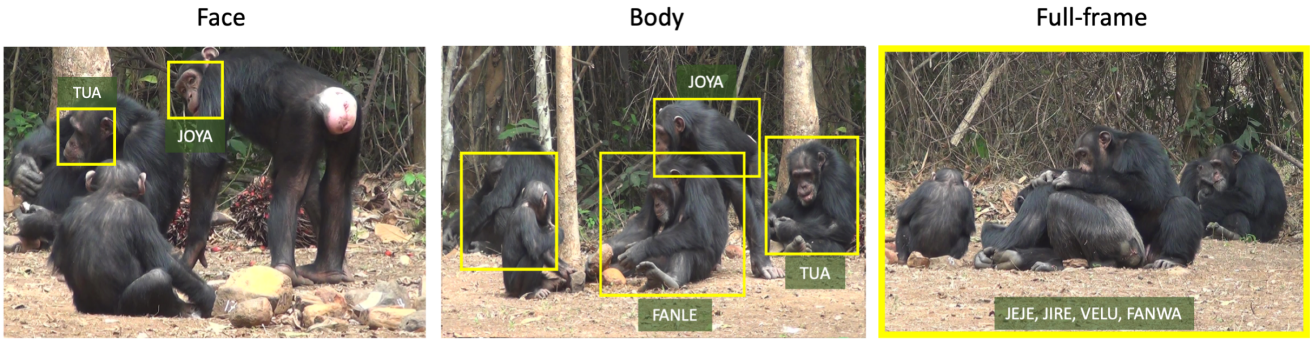


Figure 1. Levels of localisation that can be used to recognise individuals in raw footage. Left to right: (1) **Face**: high precision, but often individuals are not detected. (2) **Full Body**: while recall is higher, bodies can be incredibly difficult to detect due to their extremely deformable nature. (3) **Full Frame**: In this work we explore an architecture to recognise individuals using only frame level supervision.

make the following contributions: (i) we propose a multi-stage *Count, Crop and Recognise* (CCR) pipeline to recognise individuals from raw video with only frame level identity labels required for training. The first two *Count and Crop* stages proposes a rectangular region that tightly encloses all the individuals in the frame. The final *Recognise* stage then identifies the individuals in the frame using a multilabel classifier on the rectangular region *at full resolution* (Figure 2). (ii) We analyse the trade-offs between using our frame level model and other varying levels of localised supervision for fine-grained individual recognition (at a face, body and frame level) and their respective performances. (iii) We have annotated a large, 'in the wild' video dataset of chimpanzees with labels for multiple levels of supervision (face tracks, body tracks, frames) which is available at `TBD`. Finally, (iv) we apply a high-granularity visualisation technique to further understand the learned CNN features for the recognition of chimpanzee individuals.

2. Related Work

Animal recognition in the wild: Video data has become indispensable in the study of wild animal species [8, 43]. However, animals are difficult objects to recognise, mainly due to their deformable bodies and frequent self occlusion [1, 4]. Further, variations in lighting, other individual flora, and motion blur create additional challenges. Taking inspiration from computer-vision based systems for humans, previous methods for species identification have focused on faces, for chimpanzees [13, 21], tigers [35, 37], lemurs [12] and even pigs [25]. Compared to bodies, faces are less deformable and have a fairly standard structure. However, unlike human faces or standard non-deformable object categories, there is a dearth of readily available detectors that can be used off the shelf to localize animals in a frame, requiring researchers to annotate datasets and train their own detectors. It is also often not clear which part of the animal is the most discriminative, e.g. for elephants ears are commonly used [15], whereas for other mammals unique coat patterns such as stripes for zebras and tigers [37] and spots on Jaguars could be key for recognition [26]. Moving to a full-frame method obviates the need to identify a key discriminating region. Popular wildlife recognition datasets, such as iNaturalist [58], contain species level labels and in contrast to our dataset, typically contain a *single* instance of a class clearly visible in the foreground. While a valuable dataset does exist for the individual recognition of chimpanzees [21, 39], this dataset only contains cropped faces of individuals from zoo enclosures, less applicable to applications of conservation in the wild.

Human recognition in TV and film videos: The original paper in this area by Everingham *et al.* [18] introduced three ideas: (i) associating faces in a shot using tracking by detection, so that a face-track is the 'unit' to be labelled; (ii) the use of aligned transcripts with subtitles to provide supervisory information for character labels; and (iii) visual speaker detection to strengthen the supervision (if a person is speaking then their identity is known from the aligned transcript). Many others have adopted and extended these ideas. Cour *et al.* [11] cast the problem as one of ambiguous labelling. Subsequently, Multiple Instance Learning (MIL), was employed by [6, 27, 34, 61, 64]. Further improvements include: unsupervised and partially-supervised metric learning [9, 23]; the range of face viewpoints used (e.g. adding profile face tracks in addition to the original near-frontal face tracks) [19, 53]; and obtaining an episode wide consistent labelling [56] (by using a graph formulation and other visual cues). Recent work [42] has explored using *only* face and voice recognition, without the use of weak supervision from subtitles.

Frame level supervision: The task of labelling image regions given only frame level labels is that of weakly supervised segmentation: every image is known to have (or not) – through the image (class) labels – one or several

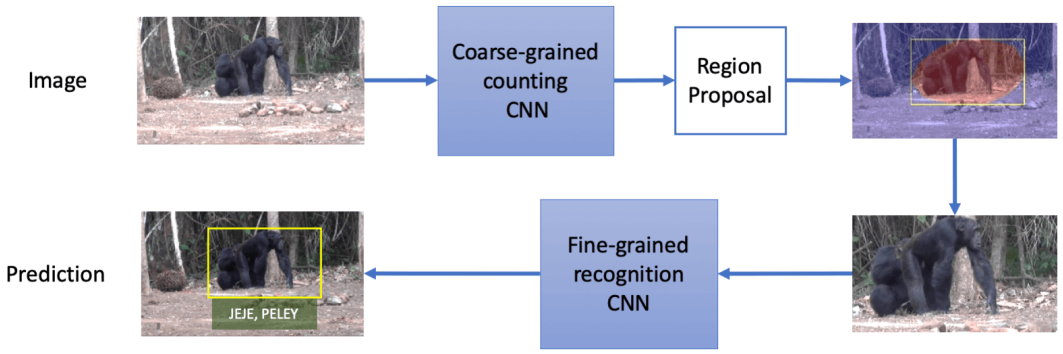


Figure 2. The *Count, Crop and Recognise* pipeline consists of three stages: (1) a coarse-grained counting network to count the number of individuals per frame, (2) a crop stage where the class activation maps from the counting network are used to localise regions of interest in the image, and (3) a fine-grained classifier trained on these cropped images.

pixels matching the label. However, the positions of these pixels are unknown, and have to be inferred. Early deep learning works on this area include [33, 46, 47]. Our problem differs in that it is fine-grained – all the object classes are chimpanzees that must be distinguished, say, rather than the 20 PASCAL VOC classes of [33, 46, 47]. While there have been works on localising fine-grained objects with weak supervision [5, 22, 29], they deal only with the restricted case of one instance per image (i.e. an image containing a single bird of class *Horned Puffin*). As far as we know, we are the first to tackle the challenging task of classifying multiple fine-grained instances in a single frame with weak supervision.

3. Count, Crop and Recognise (CCR)

Our goal is, given a frame of a video, to predict all the individuals present in that frame. We would like to learn to do this task with only *frame-level* labels, i.e no detections and hence no correspondences (who's who). The major challenge with such a method, however is that frames contain a lot of irrelevant background noise (Figure 3), and the distinctions between different individuals is often very fine-grained and subtle (these fine details are hard to learn due to the limited input resolution of CNNs).

Hence we propose a multi-stage, frame level pipeline that automatically crops discriminative regions containing individuals and so eliminates as much background information as possible, while maintaining the high resolution of the original image. This is achieved by training a deep CNN with a coarse-grained counting objective (a much easier task than fine-grained recognition), before performing identity recognition. The method is loosely inspired by the weakly-supervised object detection method C-WSL [22], however, unlike this work, our method requires neither explicit object proposals nor an existing weakly supervised detection method. Since we do not require exact bounding boxes *per instance*, but simply a generic zoomed in region, we use class guided activation maps to determine the region of focus. The multiple stages of our CCR method are described in more detail below. Precise implementation details can be found in Section 6.2, and a diagrammatic representation of the pipeline can be seen in Figure. 2.

Let $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ be a single frame of the video and let $Y \in \{0,1\}^k$ be a finite vector denoting which of the total k individuals are visible. $Y[i] = 1$ if the i-th individual is visible in $\mathbf{x}$, and $Y[i] = 0$ otherwise.

Count: We first train a parameterised function $c_\theta(\mathbf{x}')$, given a resized image input $\mathbf{x}' \in \mathbb{R}^{C\times H'\times W'}$ to count the number of individuals n within a frame. In general, we can cast this problem as either a multiclass problem or a regression problem. Since the number of individuals per frame in our datasets is small, we pose this counting task as one of multiclass classification, where the total number of individuals present can be categorised into one of the following classes $n \in \{0, 1, ..., N\}$ where all counts of N or more are binned into a single bin, with N selected as a hyperparameter (in this work we use $N = 3$). The 'Negatives' class ($n = 0$) is very important for training. Labels for counting come for free with frame level annotation (total number of labels per frame, or $n = |Y|$). The loss to be minimised can then be framed as a cross-entropy loss on the target n values. In this work we instantiate $c(\mathbf{x}')$ as a deep convolutional neural network (CNN) with convolutional layers followed by fully connected layers. Generally $H', W' < H, W$ due to the discrepancy in resolution of raw images and pretrained CNNs.

Crop: Class Activation Maps (CAMs) [67] are generated from the counting model $c_\theta(\mathbf{x}')$ to localise the discriminative regions. For resized input image $\mathbf{x}'$, let $f_k(i, j)$ denote the activation of a unit k in the last convolutional layer, and w_k^n denote the weight corresponding to count n for unit k. The CAM, M_n, at each spatial location is given by:

$$M_n(i, j) = \sum_k w_k^n f_k(i, j) \quad (1)$$

describing the importance of visual patterns at different spatial locations for a given class, in this case a count. By upsampling the CAM to the same size of $\mathbf{x}$ (H, W) image regions most relevant to the particular category can be identified. The CAM is then normalised and segmented:

$$M_n^{norm}(i,j) = \frac{M_n(i,j) - \min_{i,j} M_n(i,j)}{\max_{i,j} M_n(i,j)} \quad (2)$$

$$M_n^{thresh}(i,j) = \begin{cases} 1, & \text{if } M_n^{norm}(i,j) > T \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $T \in [0,1]$ is the chosen threshold value. The largest connected component in M_n^{thresh} is found using classical component labelling algorithms [20, 62], examples shown in Figure 3. The bounding box enclosing this component is used to crop the original input image $\mathbf{x}$ to get $\mathbf{x}_{crop}$, removing irrelevant portions of the image and permitting higher resolution of the cropping region.

Recognise: The cropped regions $\mathbf{x}_{crop}$ are used to train a fine grained recognition classifier $R_\phi(\mathbf{x}'_{crop})$ using the original frame-level labels Y. This second recognition classifier is also instantiated as a CNN, with different parameters ϕ, and trained for the task of multilabel classification, with one class for every individual in the dataset. We use a weighted Binary Cross-Entropy loss, where the weight w_i for each class i is: $w_i = f_{max}/f_i$, where f_{max} refers to the number of instances for the most populous class, and f_i is the number of instances for class i.

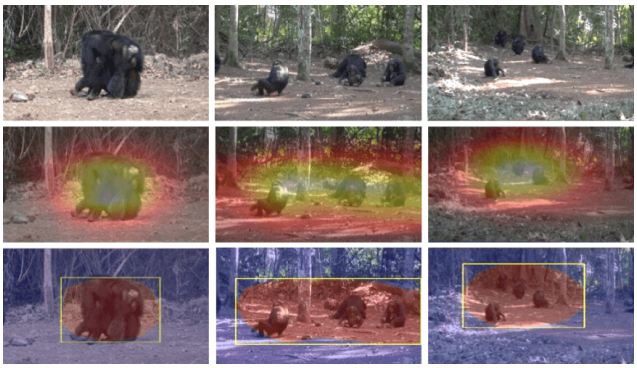


Figure 3. Region proposals for the Chimpanzee Bossou dataset. These are learnt via our counting CNN with no detection supervision at all. Top row: original frame; middle row: CAM for the count; bottom row: region proposal. Note in the second column, how the localisation works well even when the individuals are far apart from each other.

Why use counting to localise? Our method begs the following question: if a model must identify discriminative regions to be able to count individuals, surely it must also identify these regions to perform fine-grained recognition?

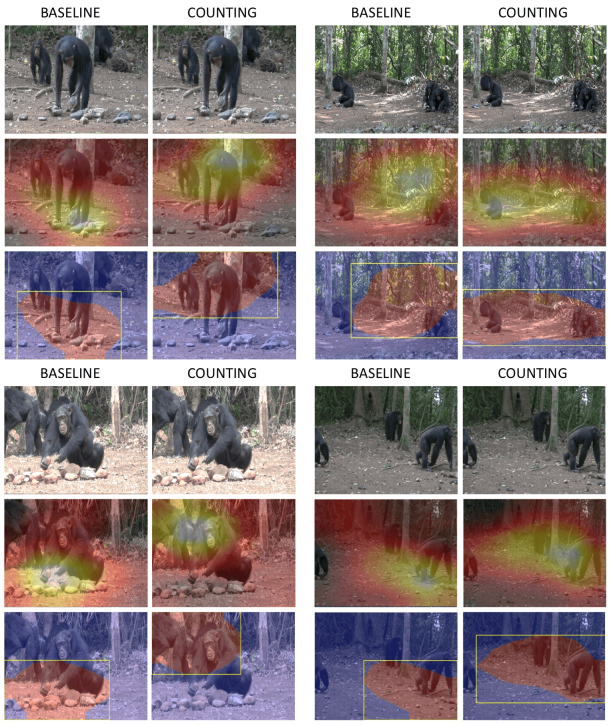


Figure 4. Region proposals for both the baseline (recognition) and counting method. Note how the baseline method mistakenly focuses on the background features, rocks and trees, to recognise individuals.

In this case we could just train the fine grained recognition network to obtain region proposals, crop regions and then retrain the recognition network in an iterative manner. However, counting objects is a much easier task than the fine-grained recognition of identities (a widely studied phenomenon in psychology, called subitizing [10] suggests that humans are able to count objects with a single glance if the total number of objects is small). We find that this leads to much better region proposals, as demonstrated in Figure 4 where we show proposals obtained from a counting model and from an identification model. By tackling an easier task first, our model is using a form of curriculum learning [3].

4. Face and Body Tracking and Recognition

In order to test recognition methods that explicitly use only face and body regions, we first create a chimpanzee face and body detection dataset, by annotating bounding boxes using the VIA annotation tool [16]. We then train a detector with these detection labels, and run the detector on every frame of the video. A tracker is then run to link up the detections to form *face-tracks* or *body-tracks*, which then become a single unit for both labelling and recognition. Examples are shown in Figure 5. Finally, we train a standard CNN multi-class classifier on the regions in the track using a cross-entropy loss on the identities in the dataset to train a recognition model.

Figure 5. Chimpanzee tracks for face (top row) and body (bottom two rows).

5. Datasets

Chimpanzee Bossou Dataset: We use a large, un-edited video corpus of chimpanzee footage collected in the Bossou forest, southeastern Guinea, West Africa. Bossou is a chimpanzee field site established by Kyoto University in 1976 [30, 41, 50, 54]. Data collection at Bossou was done using multiple cameras to document chimpanzee behaviour at a natural forest clearing (7m x 20m) located in the core of the Bossou chimpanzees' home range. The videos were recorded at different times of the day, and span a range of lighting conditions. Often there is heavy occlusion of individuals due to trees and other foliage. The individuals move around and interact freely with one another and hence faces in video have large variations in scale, motion blur and occlusion due to other individuals. Often faces appear as extreme profiles (in some cases only a single ear is visible). While the original Bossou dataset is a massive archive with over 50 hours of data from multiple years, in this paper we use roughly 10 hours of video footage from the years 2012 and 2013, of which we reserve 2 hours for testing. Chimpanzees are visible for the vast majority of this footage, therefore we also include sampled frames of just the forest background ($n = 0$) from other years to permit negative training for all methods.

Dataset annotation and statistics: We manually provide frame level annotations (i.e. name tags for the individuals present) for every frame in the videos using the VIA video annotation tool [17]. VIA is an open source project based solely on HTML, Javascript and CSS (no dependency on external libraries) and runs entirely in a web browser*. In addition, we compute face and body detections and tracks (as described in Section 4) and also label these tracks manually using the VIA tool. All identity labelling was done by an expert anthropologist familiar with the identities in the archive. The statistics of the dataset are given in Table 1. The frame-level frequency histogram for each individual is shown in Figure 6, where an instance of an individual is defined as a frame for which the individual is visible.

*http://www.robots.ox.ac.uk/~vgg/software/via/

	hours	#frames	# individuals
train	8.30	830k	13
test	1.56	161k	10
total	9.86	992k	13

Table 1. Dataset Statistics for the Chimpanzee Bossou dataset. We annotate facetracks, bodytracks and identities at a frame level.

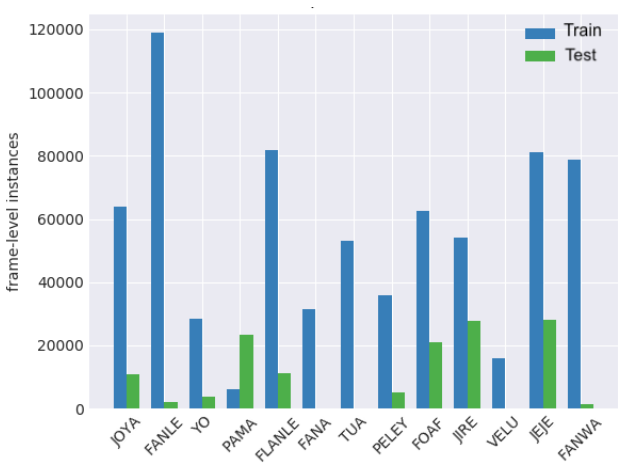


Figure 6. Instance frequency histograms for each individual in the Chimpanzee Bossou dataset.

6. Experiments

We first evaluate the performance of the face-track and body-track methods, in particular the proportion of frames that they can label (the frame recall), and their identity

recognition performance. This is then compared to the performance of the frame-level CCR method using average precision (AP) to analyse the trade-offs thoroughly. We also compare the CCR method to a simple baseline, where an identity recognition CNN is trained directly on the resized raw (not zoomed in) images $\mathbf{x}'$.

6.1. Evaluation Metrics

Detector Recall: The detector recall is the proportion of instances where faces (or bodies) are detected and tracked. This provides an upper bound on the number of individual instances that can be recognised from the video dataset using the face-track or body-track methods. We note that this is a function of two effects: (1) the visibility of the face or body in the image (faces could be turned away, be occluded etc); and (2) the performance of the detection and tracking method (i.e. is the face detected even if it is visible); though we do not distinguish these two effects here.

Identification Accuracy: This is the proportion of detections that are labelled correctly (each face-track or body-track can only be one of the possible identities).

System-level Average Precision (AP): For the face (and body) track methods, the precision and recall for each individual is computed as follows: all tracks are ranked by the score of the individual face classifier; if the track belongs to that individual, then all the frames that contain that track are counted as recalled; if the track does not belong to that individual, then the frames that contain that track are not recalled (but the precision takes these negative tracks into account), i.e. we only recall the frames containing a track if we correctly identify the individual in that track. For the frame level CCR method, the frames are ranked by the frame-level identity classifier, and the precision and recall computed for this ranked list. We then calculate both the micro and macro Average Precision score over all the individuals. Macro Average Precision (mAP) takes the mean of the AP values for every class, whereas Micro Average Precision (miAP) aggregates the contributions of all classes to compute its average metric. For our heavily class unbalanced datasets, the latter is a much better indicator of the overall performance. (histograms are provided in the supplementary material).

6.2. Implementation Details

CNN architecture and training: For a fair comparison, we use the following hyperparameters across *all* recognition models: a ResNet-18 [28] architecture pretrained on ImageNet [14] with input size $H', W' = 224$ i.e. for the counting CNN $c_\theta(\mathbf{x}')$, the fine-grained identity CNN $R_\phi(\mathbf{x}'_{crop})$, and the recognition CNNs used for both the body and the face models. This architecture achieves a good trade-off between performance and number of parameters. In principle any deep CNN architecture could be used with our method. The models are trained and tested on every third frame from the video (to avoid the large amount of redundancy in consecutive frames). We use a batch size of 64; standard data augmentation (colour jittering, horizontal flipping etc.) but only random cropping on the negative ($n = 0$) samples. All models are trained end-to-end in PyTorch [45]. Models and code will be released.

Face and Body tracks: The face and body tracks were obtained by training a Single Shot MultiBox Detector (SSD) [38], on 8k and 16k bounding box annotations respectively. The annotations were gathered on frames sampled every 10 seconds from a subset of training footage as well as from videos from other years. The detectors are trained in PyTorch with 300×300 input resolution and the same data augmentation techniques as [38]. We use a batch size of 64 and train the detectors for 95k iterations with a learning rate of $1E-4$. We used the KLT [40, 57] and SPN [36] tracker to obtain face and body tracks respectively. During the recognition stage, predictions are averaged across a track.

Count, Crop and Recognise: The coarse-grained counting CNN is applied on the entire dataset and the CAM of the highest softmax prediction for each image recorded. The CAMs, just 7×7 int arrays, are saved cheaply as grey-scale images each of size 355 bytes. Alternatively, this can be performed online during training, albeit at a greater computational cost since the CAMs are recomputed every epoch. Before training the recognition stage, we upsample the CAMs to the size of its corresponding image and threshold with $T = 0.5$, perform full-resolution cropping and then resize back to 224×224, the input size of the fine-grained identity CNN $R_\phi(x_{crop})$. Fine-grained recognition is then performed on these cropped regions.

6.3. Results

Detector recall and identification accuracy: The performance is given in Table 2. It is clear that recall is a large limitation for both the face-track and body-track methods. The face detector recall is low (less than 40%), far lower than that of the body detector. This reflects the fact that the chimpanzee's faces are not visible in many frames, rather than failures of the face detector. Hence even a perfect face recognition system would miss many chimpanzee instances at the frame level. While the identification accuracy for chimpanzees, is slightly higher for faces than for bodies, the relatively high recall of the body-track method shows a clear advantage over faces.

	#instances	#tracks	recall (%)	test acc. (%)
face	$1.02m$	$5k$	39.9	71.3
body	$1.64m$	$12k$	64.0	70.5
frame	$2.13m$	-	100.0	-

Table 2. Face and body detector recall and identification test accuracy (acc.) results for the Bossou dataset. Recall is calculated as a percentage of the total number of instances annotated at a frame level, which we note as a theoretical upper bound of 100%.

Method	mAP	miAP
Random	28.4	29.2
Face	40.1	47.1
Body	42.4	58.3
Frame Level		
Baseline	45.5	48.2
CCR	**50.0**	**59.1**

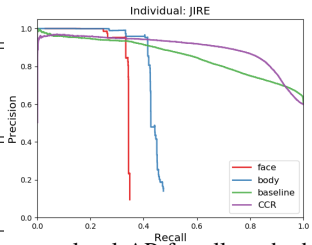


Figure 7. Left: Comparison of system level AP for all methods on the *test* set; Right: PR curves for a single individual from the Chimpanzee dataset.

System level AP: Results are given in Figure 7, left. We compare our CCR method to a simple baseline without the Count and Crop stages. CCR outperforms the baseline by a large-margin (more than 9% AP). The PR curve for the chimpanzee 'JIRE' (Figure 7, right), reiterates the results that face-track recall is the lowest, albeit with the highest precision. In contrast, the CCR method has far higher recall and with a similar level of precision. The overall AP values (Figure 7, left) show that the body-track AP is quite high, since it achieves a large boost in recall over the face-tracks with a very small drop in identification accuracy (less than 1%). We note that the CCR method, however, outperforms the body-track method as well. This is an impressive performance considering CCR requires only frame level supervision in training, and eschews the need to train a body detector.

7. Weakly Supervised Localisation of Individuals

Labelling individuals within a frame offers insight into social relationships by monitoring the frequency of co-occurrences and locations of the capturing cameras. However, unlike face and body detection, the frame level approach does not explicitly localise individuals within the frame, preventing analysis of the local proximity between individuals. To tackle this, we propose an extension to CCR which localises individuals without any extra supervisory data. This is shown in the examples of Figure 8.

Following a similar process to the 'Crop' stage in CCR, bounding boxes are generated for each labelled individual from CAMs extracted from the recognition model $R_\phi(\mathbf{x}'_{crop})$. The locations of the individuals are assumed to be at the centroid of these bounding boxes, with qualitatively impressive results even when the individuals are grouped together.

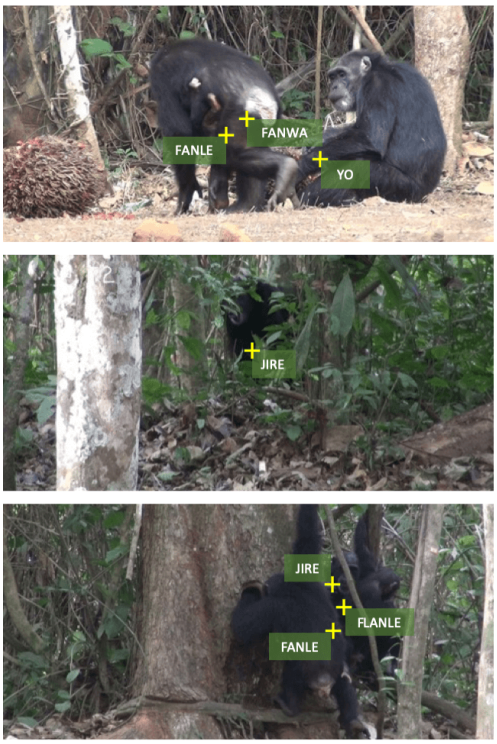


Figure 8. Weakly supervised localisation of individuals.

8. Interpretability

In this penultimate section, we introduce a high-granularity visualisation tool to understand and interpret the predictions made by the face and body recognition models. These tease out the discriminative features learnt by the model for this task of fine-grained recognition of individuals. Understanding these features can provide new insights to human researchers.

A Class Activation Map (CAM) [67], introduced in Section 3, can be used to localise discrimnative features but it does so at low resolution and thus cannot identify high-frequency features, such as edges and dots. An alternative visualisaton method is Excitation Backprop (EBP) [66]. EBP achieves high-granularity visualisation via a top-down attention model, working its way down from the last layer of the CNN to the high resolution input layer. Activations are followed from layer to layer with a probabilistic Winner-Take-All process.

In Figure 9, we show the EBP visualisations from the face recognition model of example images of individuals in the Bossou dataset. When the ears are visible, the face model shows high activation on the ear region – similarly for the brow and mouth regions. Upon closer inspection of the original face images, the ears of each individual are in-

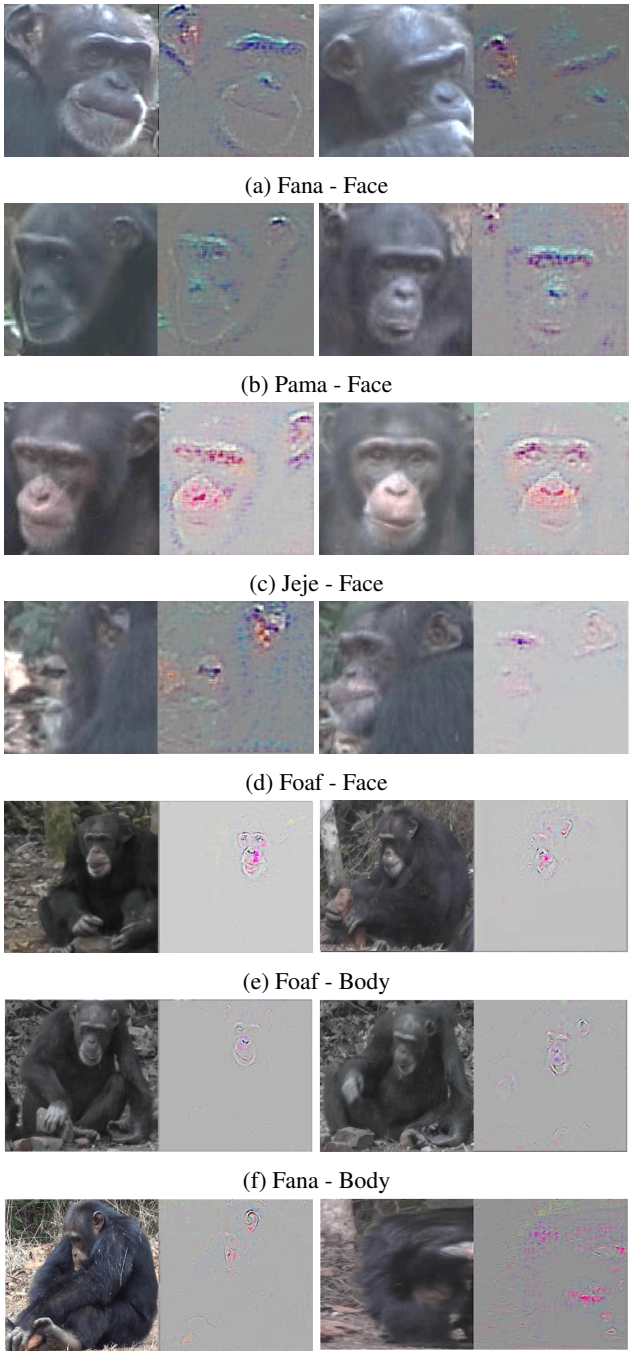
deed highly unique and distinguishable. The expert anthropologist, who manually labelled the dataset, noted that he doesn't pay particular attention to the ears when identifying the individuals. Perhaps our discovery of ear uniqueness in chimpanzees in this dataset, and possibly all chimpanzees, could improve expert's recognition of chimpanzee individuals.

The EBP visualisation for the body recognition model in Figure 9 reiterates the importance of the face and ears in distinguishing the individuals. Further, note Jejes hairless patch on his left leg in the top of Figure 9g and corresponding EBP activation, indicating that the body recognition model also uses distinguishing marks on the body. Similarly, Foafs white spot above his upper lip (Figure 9e) is another region of high activation. The presence of the white spot was unbeknownst to the anthropologist who noted he would now use this information to identify Foaf in the future. These two examples show that a CNNs learned discriminative features for a specific individual can be visualised and interpreted by humans. Of course, these findings are not statistically relevant and quantitative analysis would be needed in order to determine the effectiveness of the use of recognition CNNs to train human experts.

9. Conclusion

We have proposed and implemented a simple pipeline for fine-grained recognition of individuals using only frame-level supervision. This has shown that a counting objective allows us to learn very good region proposals, and zooming into these discriminative regions gives substantial gains in recognition performance. Many datasets 'in the wild' have the property that resolution of individuals can vary greatly with scene depth, and with cameras panning and zooming in and out. Our frame-level method approaches the precision of face-track and body-track recognition methods, whilst now allowing a much higher recall. We hope that our newly created dataset will spur further work in high-recall frame-level methods for fine-grained individual recognition in video, and that our preliminary work on interpretability of CNNs for classifying individuals of species gives insight on identifying discriminative features.

Acknowledgments: This project has benefited enormously from discussions with Dora Biro and Susana Carvalho at Oxford. We are grateful to Kyoto University's Primate Research Institute for leading the Bossou Archive Project, and supporting the research presented here, and to IREB and DNRST of the Republic of Guinea. This work is supported by the EPSRC programme grant Seebibyte EP/M013774/1. A.N. is funded by a Google PhD fellowship; D.S. is funded by the Clarendon Fund, Boise Trust; Fund and Wolfson College, University of Oxford.



(a) Fana - Face

(b) Pama - Face

(c) Jeje - Face

(d) Foaf - Face

(e) Foaf - Body

(f) Fana - Body

(g) Jeje - Body

Figure 9. Excitation Backprop [66] visualisations (right) from the face and body recognition model for example images of of individuals in the Chimpanzee Bossou dataset.

We also thank Dr Ernesto Coto, his assistance was paramount to the success of this work.

References

[1] H. M. Afkham, A. T. Targhi, J.-O. Eklundh, and A. Pronobis. Joint visual vocabulary for animal classification. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 2

[2] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv preprint arXiv:1611.01484*, 2016. 1

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 4

[4] T. L. Berg and D. A. Forsyth. Animals on the web. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1463–1470. IEEE, 2006. 2

[5] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 3

[6] P. Bojanowski, F. Bach, , I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proc. ICCV*, 2013. 1, 2

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018. 1

[8] A. Caravaggi, P. B. Banks, A. C. Burton, C. M. Finlay, P. M. Haswell, M. W. Hayward, M. J. Rowcliffe, and M. D. Wood. A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3):109–122, 2017. 2

[9] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *Proc. ICCV*, pages 1559–1566, 2011. 2

[10] D. H. Clements. Subitizing: What is it? why teach it? *Teaching children mathematics*, 5:400–405, 1999. 4

[11] T. Cour, B. Sapp, and B. Taskar. Learning from ambiguously labeled images. In *Proc. CVPR*, 2009. 1, 2

[12] D. Crouse, R. L. Jacobs, Z. Richardson, S. Klum, A. Jain, A. L. Baden, and S. R. Tecot. Lemurfaceid: a face recognition system to facilitate individual identification of lemurs. *Bmc Zoology*, 2(1):2, 2017. 2

[13] D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tymoszek, and A. Jain. Face recognition: Primates in the wild. *arXiv preprint arXiv:1804.08790*, 2018. 1, 2

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 6

[15] I. Douglas-Hamilton. *On the ecology and behaviour of the African elephant*. PhD thesis, University of Oxford, 1972. 2

[16] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/~vgg/software/via/, 2016. 4

[17] A. Dutta and A. Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. 5

[18] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proc. BMVC*, 2006. 1, 2

[19] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009. 2

[20] C. Fiorio and J. Gustedt. Two linear time union-find strategies for image processing. *Theoretical Computer Science*, 154(2):165 – 181, 1996. 4

[21] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kühl, and J. Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition*, pages 51–63. Springer, 2016. 2

[22] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168, 2018. 3

[23] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *Proc. ECCV*, pages 634–647, 2010. 2

[24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic Imaging*, 2016(11):1–6, 2016. 1

[25] M. F. Hansen, M. L. Smith, L. N. Smith, M. G. Salter, E. M. Baxter, M. Farish, and B. Grieve. Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98:145–152, 2018. 2

[26] B. J. Harmsen, R. J. Foster, E. Sanchez, C. E. Gutierrez-González, S. C. Silver, L. E. Ostro, M. J. Kelly, E. Kay, and H. Quigley. Long term monitoring of jaguars in the cockscomb basin wildlife sanctuary, belize; implications for camera trap studies of carnivores. *PloS one*, 12(6):e0179505, 2017. 2

[27] M. Haurilet, M. Tapaswi, Z. Al-Halah, and R. Stiefelhagen. Naming tv characters by watching and analyzing dialogs. In *WACV*, 2016. 1, 2

[28] 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*, pages 770–778, 2016. 6

[29] T. Hu and H. Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint*

arXiv:1901.09891, 2019. 3

[30] T. Humle. Location and ecology. In *The chimpanzees of Bossou and Nimba*, pages 13–21. Springer, 2011. 5

[31] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst Technical Report, 2010. 1

[32] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 1

[33] A. Kolesnikov and C. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. ECCV*, 2016. 3

[34] M. Köstinger, P. Wohlhart, P. Roth, and H. Bischof. Learning to recognize faces from videos and weakly relatedinformation cues. In *avss*, 2011. 1, 2

[35] H. S. Kühl and T. Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in ecology & evolution*, 28(7):432–441, 2013. 2

[36] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[37] S. Li, J. Li, W. Lin, and H. Tang. Amur tiger re-identification in the wild. *CoRR*, abs/1906.05586, 2019. 2

[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. 6

[39] A. Loos and A. Ernst. Detection and identification of chimpanzee faces in the wild. In *2012 IEEE International Symposium on Multimedia*, pages 116–119. IEEE, 2012. 2

[40] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 6

[41] T. Matsuzawa. Field experiments of tool-use. In *The Chimpanzees of Bossou and Nimba*, pages 157–164. Springer, 2011. 5

[42] A. Nagrani and A. Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *Proc. BMVC*, 2017. 1, 2

[43] T. Nishida, K. Zamma, T. Matsusaka, A. Inaba, and W. C. McGrew. *Chimpanzee behavior in the wild: an audio-visual encyclopedia*. Springer Science & Business Media, 2010. 2

[44] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. BMVC*, 2015. 1

[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[46] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[47] P. H. O. Pinheiro and R. Collobert. Weakly supervised semantic segmentation with convolutional networks. In *Proc. CVPR*, 2015. 3

[48] H. Rakotonirina, P. M. Kappeler, and C. Fichtel. The role of facial pattern variation for species recognition in red-fronted lemurs (eulemur rufifrons). *BMC evolutionary biology*, 18(1):19, 2018. 1

[49] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2012. 1

[50] D. Schofield, A. Nagrani, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5, 2019. 5

[51] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[52] S. Sinha, M. Agarwal, M. Vatsa, R. Singh, and S. Anand. Exploring bias in primate face detection and recognition. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 541–555, Cham, 2019. Springer International Publishing. 1

[53] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *Proc. CVPR*, 2009. 2

[54] Y. Sugiyama. Population dynamics of wild chimpanzees at bossou, guinea, between 1976 and 1983. *Primates*, 25(4):391–400, 1984. 5

[55] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 1

[56] M. Tapaswi, M. Baeuml, and R. Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv series. In *Proc. CVPR*, 2012. 1, 2

[57] C. Tomasi and T. K. Detection. Tracking of point features. Technical report, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, 1991. 6

[58] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2

[59] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. Iarpa janus benchmark-b face

dataset. In *CVPR Workshop on Biometrics*, 2017. 1

[60] C. L. Witham. Automated face recognition of rhesus macaques. *Journal of neuroscience methods*, 300:157–165, 2018. 1

[61] P. Wohlhart, M. Köstinger, P. M. Roth, and H. Bischof. Multiple instance boosting for face recognition in videos. In *DAGM-Symposium*, 2011. 1, 2

[62] K. Wu, E. Otoo, and A. Shoshani. Optimizing connected component labeling algorithms. volume 5747, 04 2005. 4

[63] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014. 1

[64] J. Yang, R. Yan, and A. G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *ACM Multimedia*, 2005. 1, 2

[65] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 1

[66] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. *CoRR*, abs/1608.00507, 2016. 7, 8

[67] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3, 7