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

Max Bain, Arsha Nagrani, Daniel Schofield, Andrew Zisserman

Visual Geometry Group, University of Oxford

UNIVERSITY OF OXFORD

VGG

UNIVERSITY OF OXFORD

Recognising Animal Individuals in a Video

- The aim: label animal individuals in every frame of a video

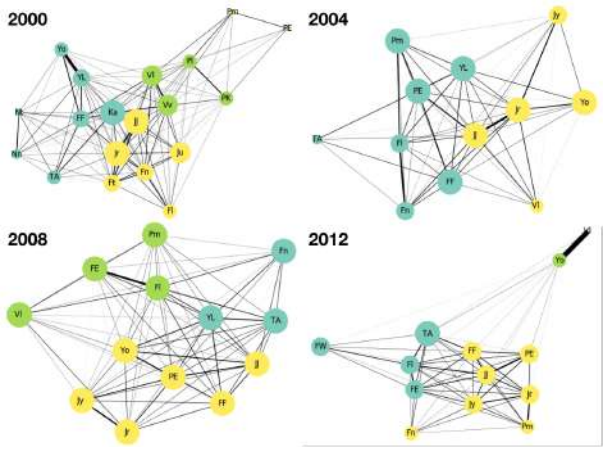
King Kong
(King Kong 2005)

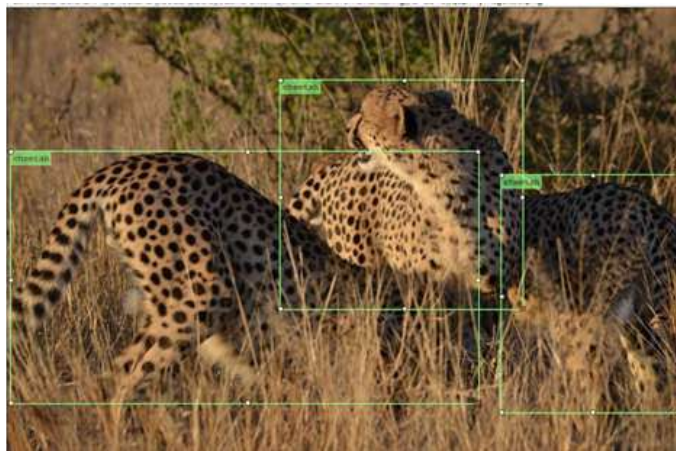
George
(Rampage 2018)

Jambo
(Durell Wildlife Park, France)

Recognising Animal Individuals in a Video

- The aim: label animal individuals in every frame of a video





Recognising Animal Individuals in a Video

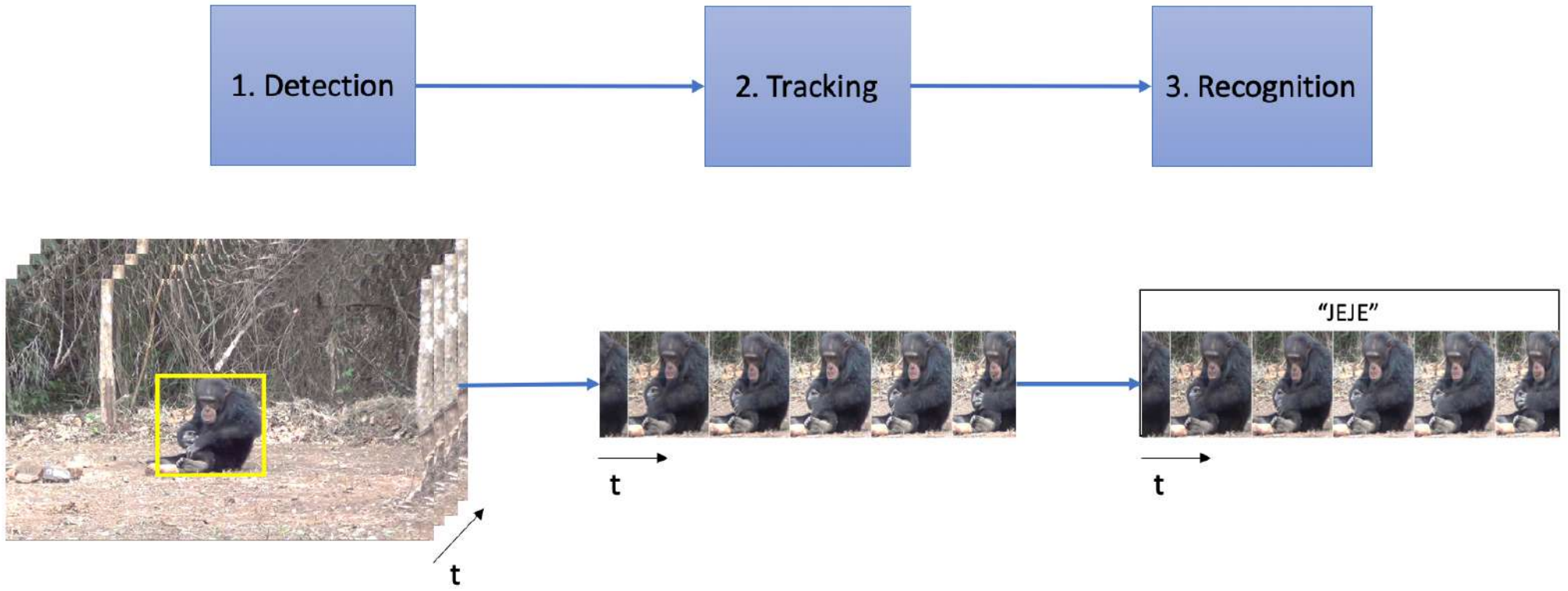
- The aim: label animal individuals in every frame of a video

Current Methods

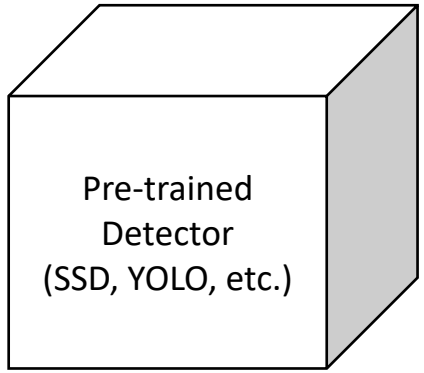
Chimpanzee face recognition from videos in the wild using deep learning
D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, S. Carvalho
Science Advances, 2019
www.robots.ox.ac.uk/~vgg/research/ChimpanzeeFaces

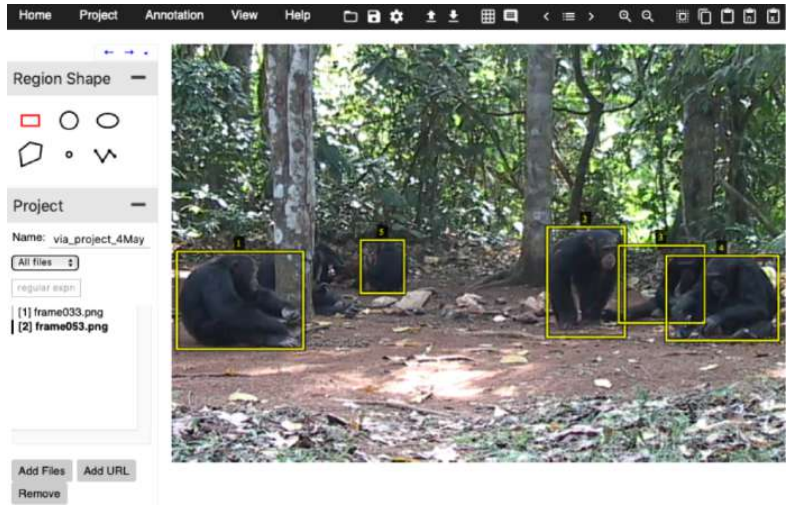
Current Methods

Current Methods: Training Pipeline

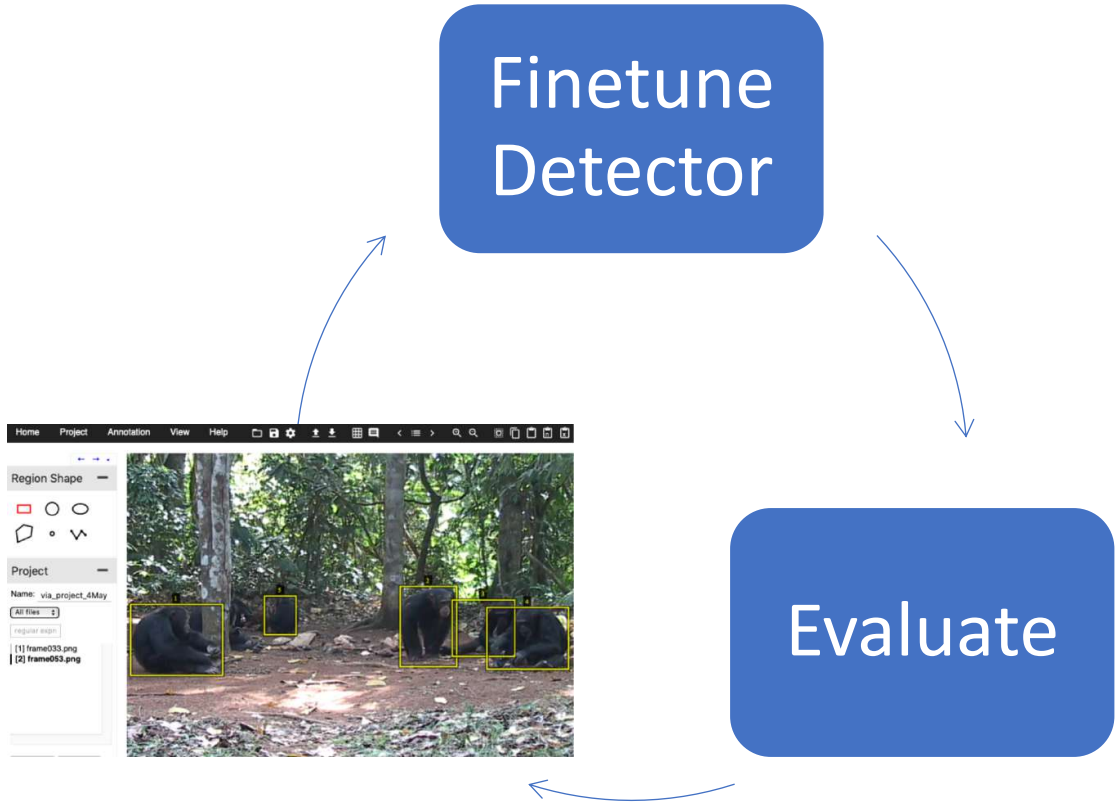


Current Methods: Detection

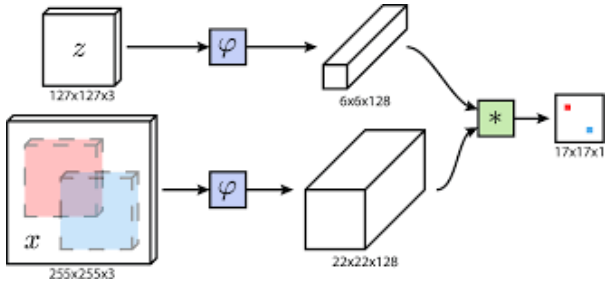


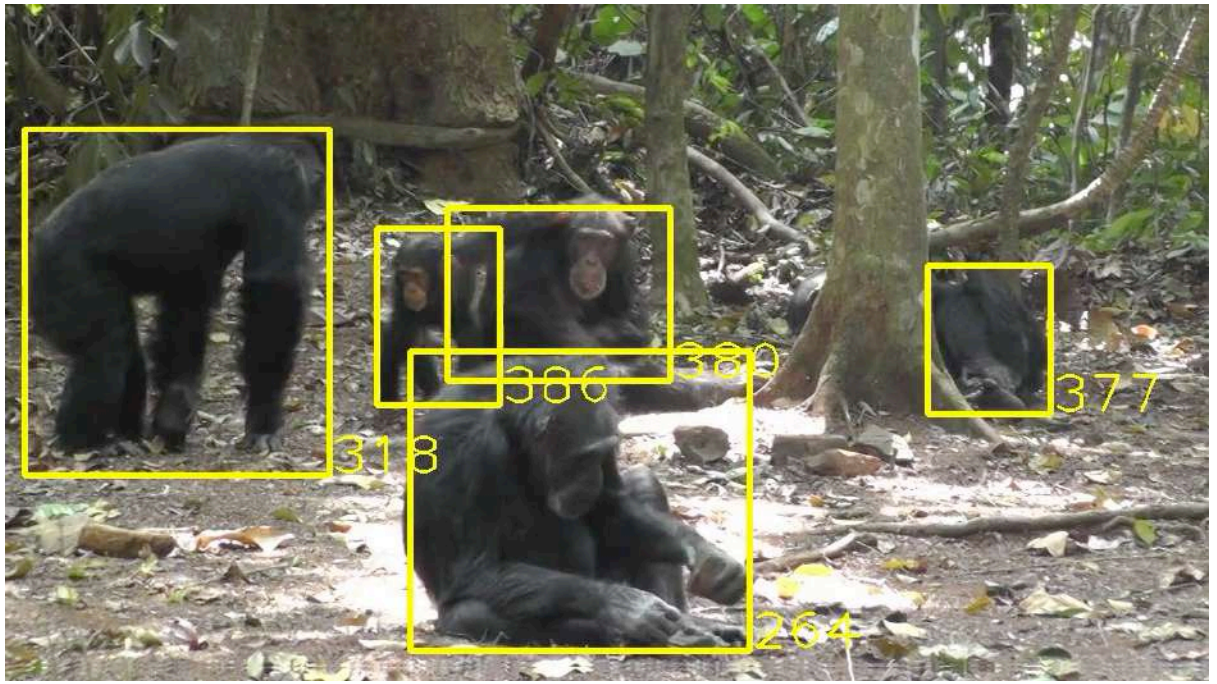


Current Methods: Detection



Current Methods: Tracking

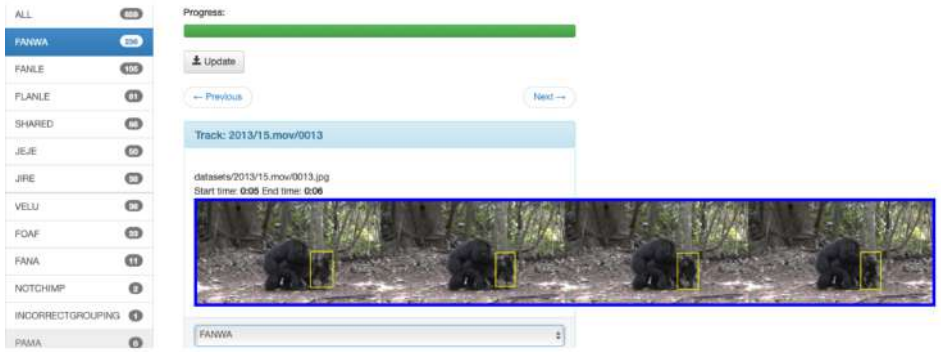




Fully-Convolutional Siamese Networks for Object Tracking
L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. S. Torr
CVPR 2017

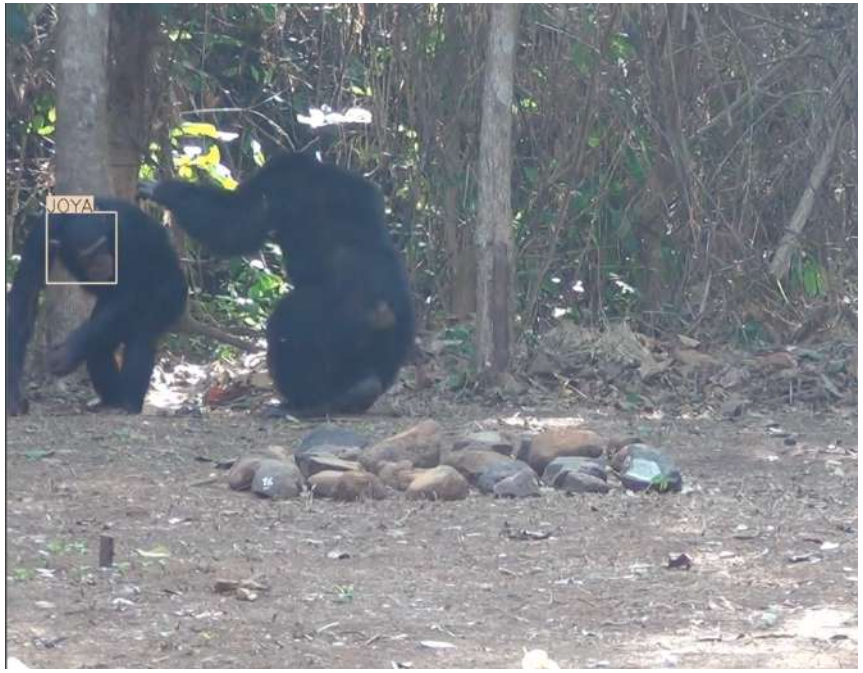
Current Methods: Recognition

1. Acquire identity labels from expert



2. Train Identity Recognition CNN
   - ResNet
   - CE weighted loss (class imbalance)

Current Methods: Challenges



Face often turned away

Bodies prone to heavy occlusion and overlap

Current Methods: Challenges

Face often turned away

Bodies prone to heavy occlusion and overlap

Current Methods: Challenges

Lack of contextual information

Current Methods: Challenges

Lack of contextual information

Current Methods: Challenges

Lack of contextual information

What if we recognised without explicit detection?

Dataset (Publicly Available)

13 Individuals

Identity
Gender
Age

10 Hours of video footage

over 2 years

1.0M Face detections w. labels

1.6M Body detections w. labels

2.1M Frame-level instances

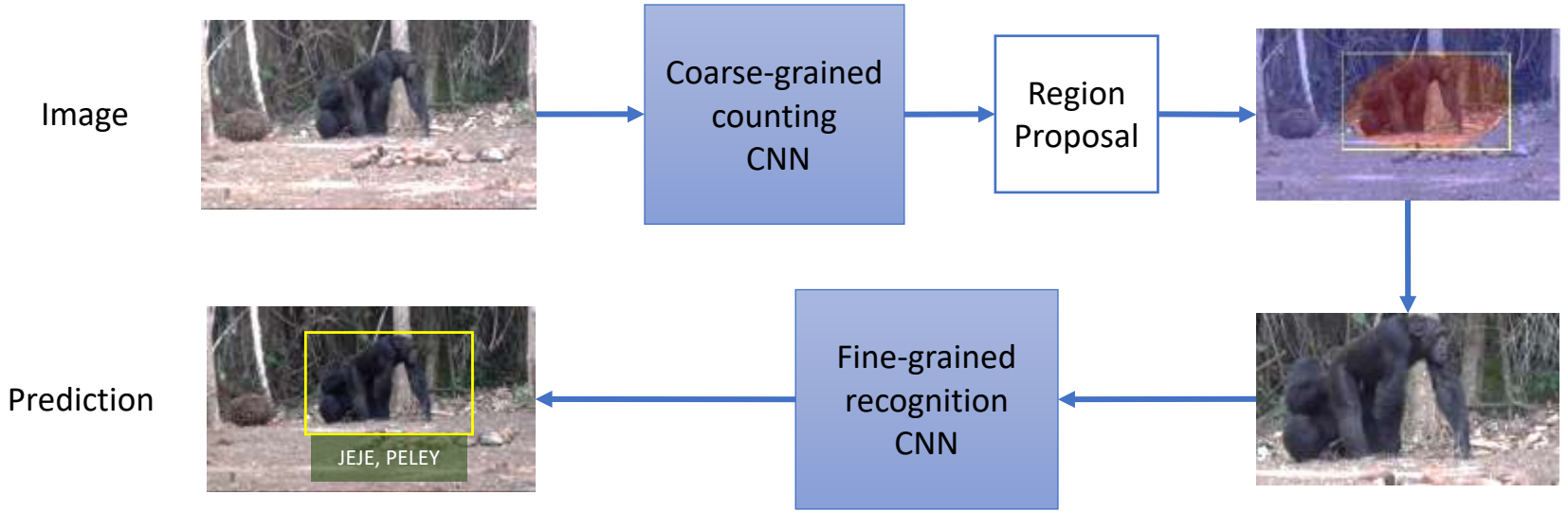
Located in a wild forest

Guinea
Africa

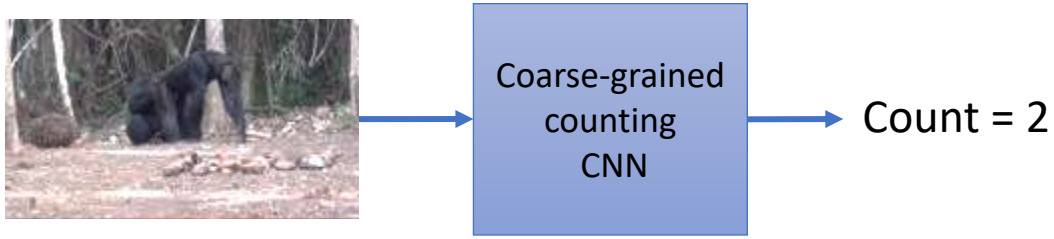
Frame-Level Recognition

- Mutli-label classifier on raw frames
- ResNet18, Sigmoid + Weighted BCE loss

Count, Crop and Recognise (CCR)

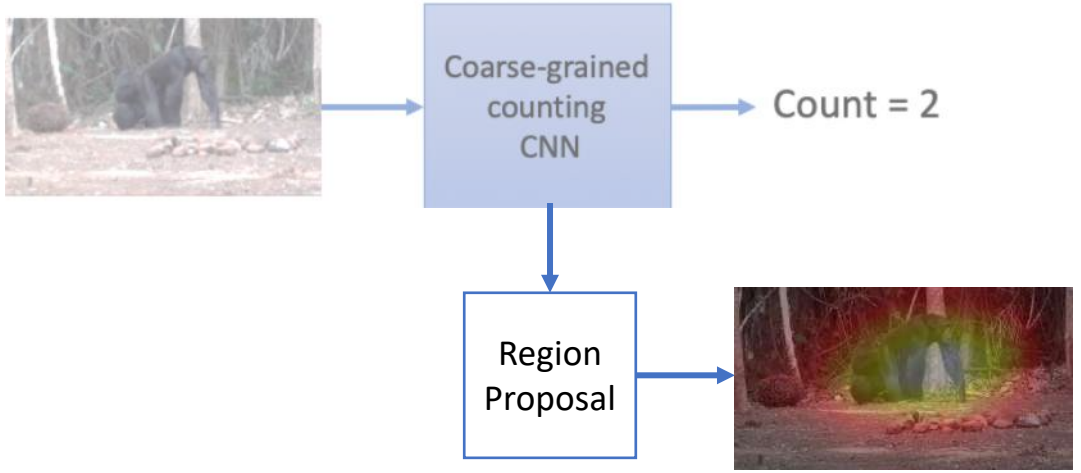


Count, Crop and Recognise (CCR)

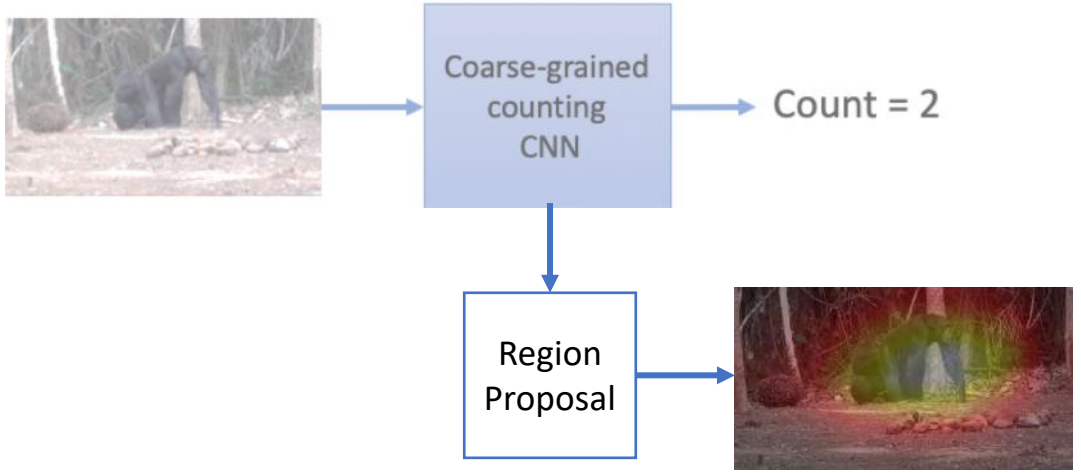


- Count labels are for free
- Trained as classification task (ResNet18, CE loss)
- Bin count of N+ into the same class

Count, **Crop** and Recognise (CCR)

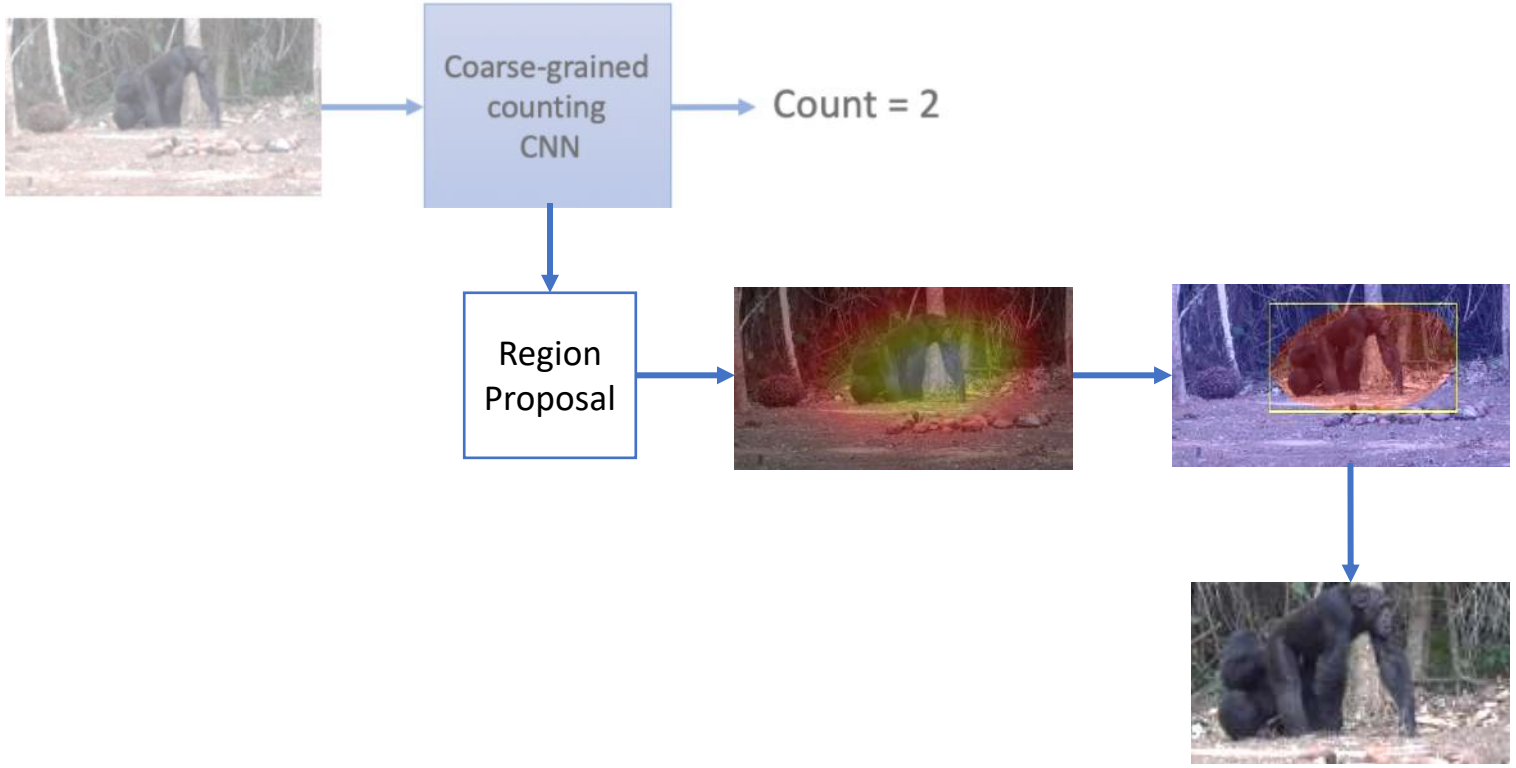


Count, **Crop** and Recognise (CCR)

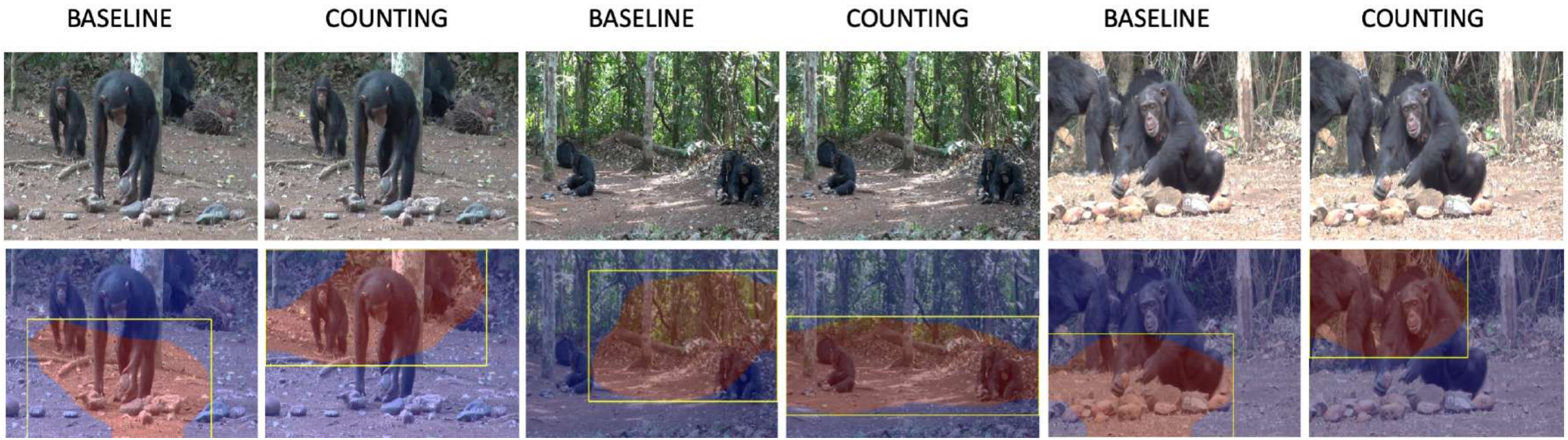


B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR'16 (arXiv:1512.04150, 2015).

Count, **Crop** and Recognise (CCR)

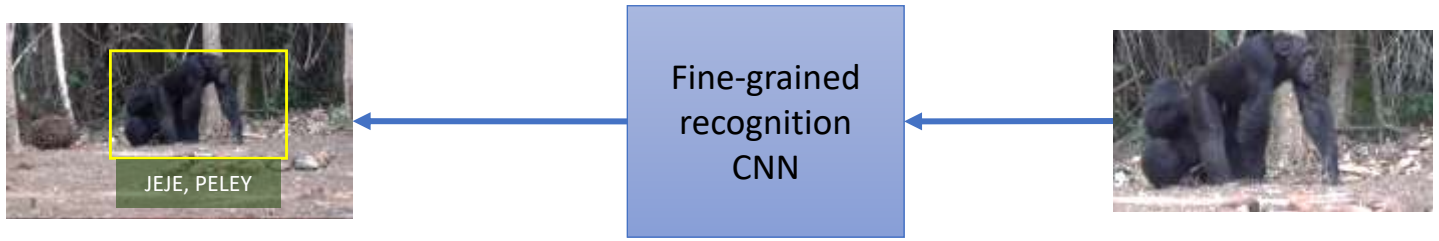


Results



Count, Crop and **Recognise** (CCR)

- Recognise



  - Multi-label classification (ResNet18)
  - Sigmoid + BCE Loss

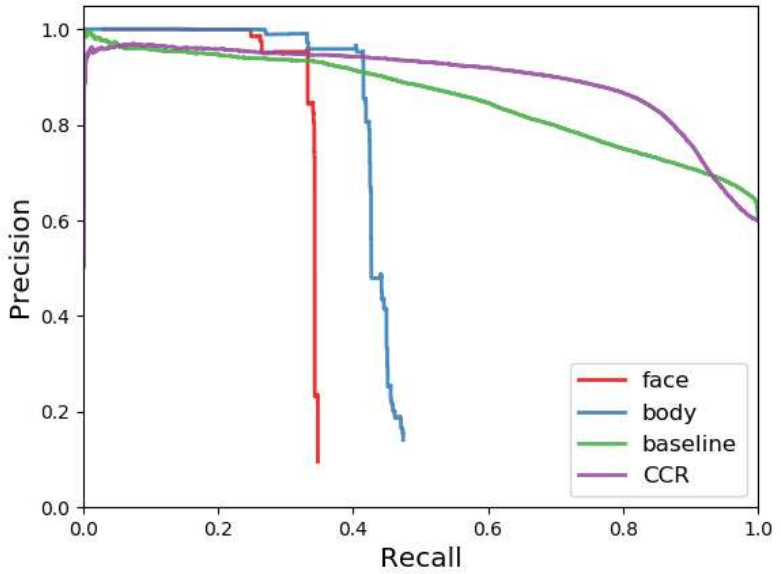
Results

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

Results

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

Individual: JIRE

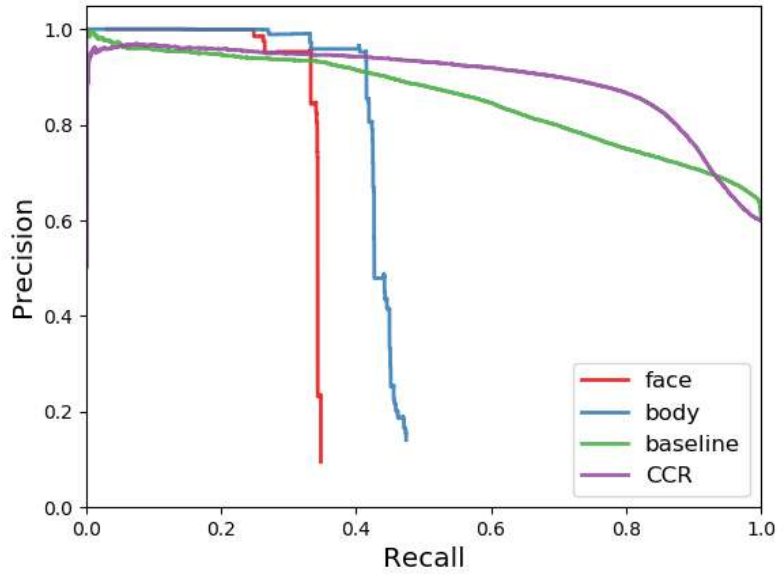


Method	AP
Face	31.2
Body	42.3
Baseline	82.3
CCR	86.4

Results

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

Individual: JIRE



Method	AP
Face	31.2
Body	42.3
Baseline	82.3
CCR	86.4

	#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	-

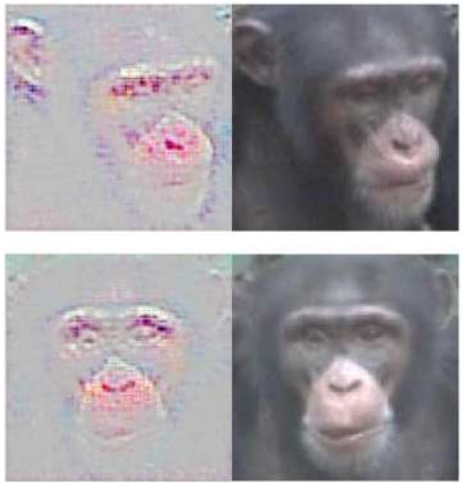
Conclusions

- Detect, Track and Recognise pipelines limited by detector performance
- Body > Face
- Frame-level recognition offers an alternative, more research needed

An quirky post-hoc application

FANA

JEJE



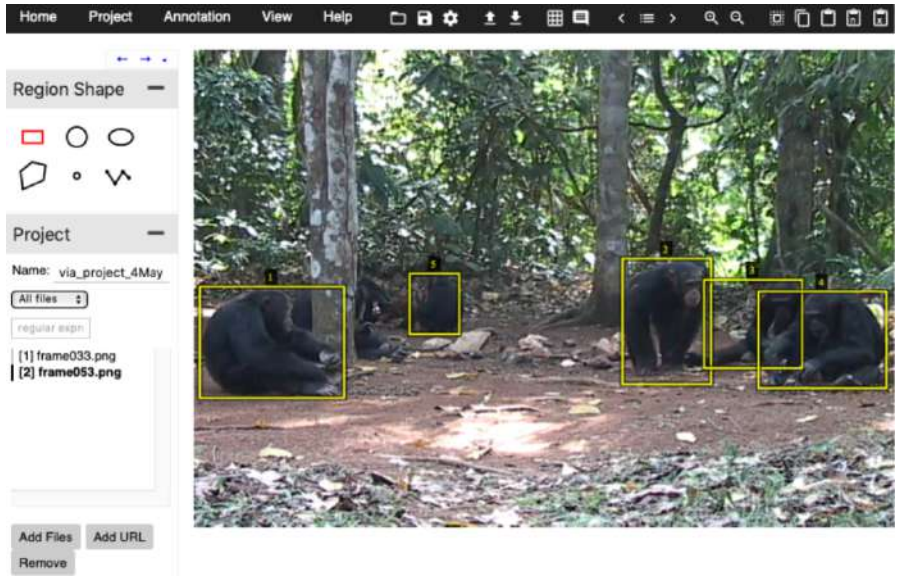
PAMA

FOAF

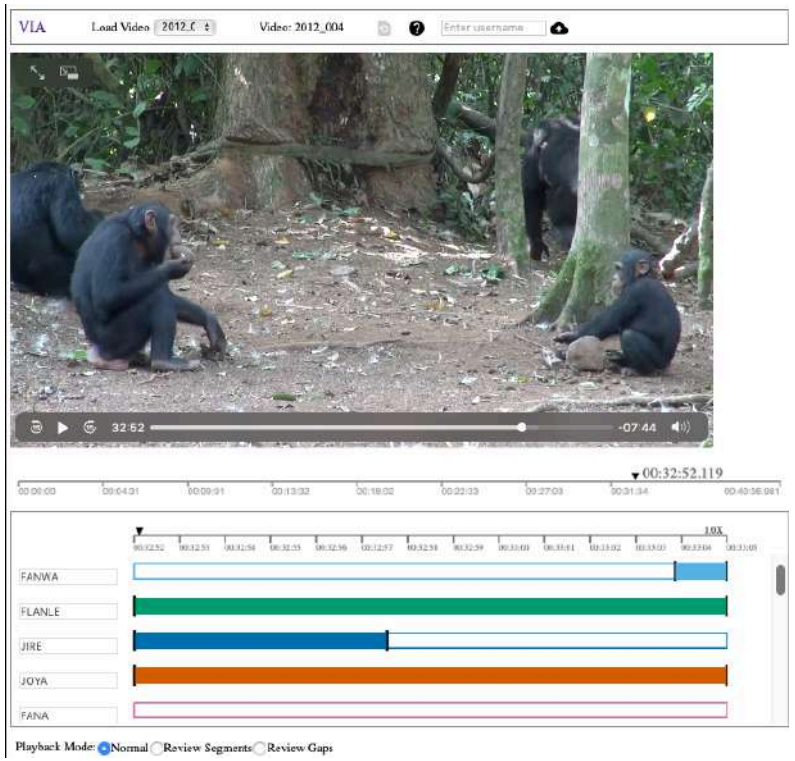
Top-Down Neural Attention by Excitation Backprop, ECCV 2016
J. Zhang, Z. Lin, J. Brandt, X. Shen and S. Sclaroff

Annotation Tools

Object Annotator (bounding boxes, keypoints, pose)



Temporal segmentation (presence, actions, speech)



Abhishek Dutta and Andrew Zisserman. 2019.

The VIA Annotation Software for Images, Audio and Video.

In Proceedings of the 27th ACM International Conference on Multimedia (MM '19)

Paper, Dataset and Code at:
www.robots.ox.ac.uk/~vgg/research/ccr

Thank you to the
organisers for arranging
this workshop!

UNIVERSITY OF
OXFORD

VGG
UNIVERSITY OF OXFORD