Applied Animal Behaviour Science 255 (2022) 105708

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim

A deep transfer learning model for head pose estimation in rhesus macaques during cognitive tasks: Towards a nonrestraint noninvasive 3Rs approach

Emily J. Bethell[a,*,1], Wasiq Khan[b,*,1], Abir Hussain[b,c]

[a] School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool L3 3AF, UK
[b] School of Computer Science & Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK
[c] College of Engineering, University of Sharjah, United Arab Emirates

ARTICLE INFO

Keywords:
3Rs
Animal behavior
Animal welfare
Attention bias
Deep gaze analysis
Deep learning
Head pose estimation
Macaca mulatta
Social threat

ABSTRACT

Head orientation is a measure of attention used in behavioral psychological research with non-human primates. It is used across a broad range of disciplines and settings, from the field to the laboratory. Field methods are time consuming with risk of coding bias and visibility issues with free-ranging animals. Laboratory methods may require restraint and use of invasive procedures. Automated systems to measure head orientation in unrestrained animals, that are robust to partial occlusion of the head, would improve coding efficiency and accuracy and provide 3Rs animal welfare benefits. We present a free-to-use deep transfer learning model for non-invasive head pose estimation in unrestrained *Macaca mulatta* taking part in cognitive experiments. Monkeys housed in social groups were filmed viewing two conspecific face stimuli presented on either side of a video camera. Video frames were manually annotated for three head positions relative to the video camera: ‘left’, ‘center’ and ‘right’. The dataset (total = 8135 images from 26 monkeys) was partitioned into training and testing datasets using a leave-k-out strategy, so that 70% of the images were used in training and 30% were used in testing. We used the VGG16, VGG19, InceptionV3 and Resnet50 as base models to train the proposed head pose classifier. We achieved model accuracy up to 93 %. The head pose estimation model presented here will be of use across contexts ranging from field-based playback experiments to assessment of welfare in zoo and clinical veterinary settings and refinement of neuroscience research practices. Model code with instructions is provided.

1. Introduction

Head orientation towards stimuli (such as visual images presented on a screen, or the direction of a sound) is a widely used response measure of attention in behavioral cognitive and neuropsychological research with human and non-human primates (Adade and Das, 2019; Pfefferle et al., 2014; Wilson et al., 2020). Social diurnal primates including humans have specialized brain areas sensitive to visual gaze cues such as head orientation, supporting the biological value of this social measure of attention (Deaner and Platt, 2003; Hadjidimitrakis, 2020; Taubert et al., 2020; Wilson et al., 2000). In real-world settings where individuals can move freely, head orientation is more easily and accurately detected than other measures such as eye-gaze. Due to the coordinated movement of the eyes and head such that the head typically aligns with the direction of eye gaze, in unrestrained contexts head orientation provides a reliable proxy measure for the direction of visual attention (Hadjidimitrakis, 2020; Itti et al., 2003).

Where head orientation is measured in naturalistic or free ranging settings, head orientation towards stimuli is typically filmed for manual coding, allowing for accuracy checks and later reliability testing. Examples include field playback experiments using auditory cues (Pfefferle et al., 2014), looking time paradigms using visual stimuli (Mandalaywala et al., 2014; Winters et al., 2015), and gaze following paradigms (Ferrari et al., 2000; Ghazanfar and Santos, 2004), all of which recorded video material. Direction of head turn relative to stimuli (left/right) is used as an indicator of hemispheric specialization in

* Corresponding authors.
E-mail addresses: E.J.Bethell@ljmu.ac.uk (E.J. Bethell), W.Khan@ljmu.ac.uk (W. Khan).

[1] Joint first authors

https://doi.org/10.1016/j.applanim.2022.105708
Received 8 December 2021; Received in revised form 8 July 2022; Accepted 12 July 2022
Available online 16 July 2022
0168-1591/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

information processing (Rogers, 2010; Teufel et al., 2007). Manual coding of video is time consuming requiring initial coder training, reliability assessment, time to sort and annotate video, and is at risk of bias where experiments are not double blind.

In the laboratory, measures of primate attention can be more highly controlled. Head orientation is an essential component of the gaze response and where eye-tracking devices are used is typically controlled mechanically, most often by fixing the subject's head in place using surgical implantation of a headpost fixing: (Adade and Das, 2019; Adams et al., 2007; Wilson et al., 2020). Mechanical restraint of head movement raises ethical issues and poses challenges for quality of science (Prescott and Lidster, 2017), and does not allow for a full understanding of natural (whole body) responses during testing (Berger et al., 2020). Currently, there are no established methods to track eye gaze in freely moving primates that are both noninvasive and non-restraint. This is partly due to a lack of available resources that coordinate head orientation with eye movement data from non-human primates (Hopper et al., 2020). Software that reliably tracks head orientation from video of unrestrained primates is a necessary first step to new improved approaches to measuring eye gaze in unrestrained animals (Hopper et al., 2020).

Recent progress in computer vision science, particularly in the field of deep learning (DL) and convolutional neural networks (CNN), indicates this is a fertile area for development of a new range of tools for noninvasive behavioral assessment (Khan et al., 2020; LeCun et al., 2015). Computer vision is a field of computer science that deals with the extraction and processing of information from digital images such as video. DL is a type of machine learning which involves a large number of layers and 'neurons' to process big data, and CNN are node-based neural networks used specifically in computer vision DL applications (i.e. with digital images). Most applications of DL and CNN within the behavioral sciences to date have been conducted with humans (Belhadi et al., 2021; Bhouri, 2021; Huang, 2021), and a number of platforms offering DL and computer vision-based tools for assessment of human head pose estimation and gaze have emerged in recent years.

Developments in head pose estimation indicate that DL approaches may provide benefits over other ML approaches. For example, Bailly and Milgram (2009) introduced a feature selection approach using fuzzy functional criteria along with boosting over generalised regression neural networks for the head pose estimation. Similarly, Wang and Song (2014) presented a multi-stage supervised manifest learning approach for human head pose estimation. In both cases, the proposed model achieved high accuracy compared to similar methods when tested over standard datasets and varying illuminations. However, in both studies, the analyses lacked evidence of statistical significance when compared with DL models such as CNN. Li et al. (2020) applied CNN for head pose estimation in people in both indoor and outdoor settings allowing image processing for head pose that was independent of landmark identification tools. Yin et al. (2017) proposed a deep 3D Morphable Model and face recognition CNNs for the classification of large pose variations in unconstrained environments with the ability to expand the pose ranges to 90∘. McCay et al. (2020) used DL for the detection of abnormal infant movements for early diagnosis of cerebral palsy from video sequences; ultimately, they retained only joint movements and excluded head pose from the model '*due to self occlusion*', indicating the model was not robust to occlusion caused by natural face-directed behavior in unrestrained infants. Khan et al. (2021) applied computer vision and machine learning to extract micro-features including facial movements, head pose and gaze information from image frames, that allowed classification of truthful and deceptive behaviors by human participants taking part in a truthful/deceptive role play. However, they used parallel conventional machine learning (ML) algorithms to extract the real-time object localizations (i.e., head pose, facial and eye movements) which could be simplified with DL algorithms for automated feature extraction for such tasks. More specifically, deep transfer learning (DTL) utilizes pre-trained models for application across contexts, thereby improving processing efficiency, generalization, and saving time traditional approaches need for initial training. To our knowledge, deep transfer learning has not yet been applied for head pose estimation in humans, despite evidence it could provide a number of benefits over existing approaches. Deep learning tools for pose estimation and behavior analysis in non-human animals are beginning to emerge but are still in their infancy (Nath et al., 2019; Valletta et al., 2017). At the

Table 1
Summary data for the 26 adult females whose video was used in the raining and testing of the DTL- HPE model. Video#: The twenty six videos of the first experimental trial was selected at random from a database of 108 animals; Threat face location: indicates whether the threat face was in the left or right location within the apparatus (and therefore presented to the left or right visual field: LVF and RVF respectively). Age: all monkeys were sexually mature adults; Social group: Monkeys were housed in social breeding groups of varying sizes containing on breeding male (n indicates no breeding male present at the time video was collected). Group size: the number of adults (excludes sexually immature individuals). Total frames: the total number of image frames extracted from the video for that individual.

Video#	Threat face location	Age (yrs)	Social group	Group size	Total Frames
1	LVF	8	n	10	406
2		11	so	15	427
3		12	th	7	462
4		12	so	9	254
5		12	so	13	239
6		12	so	12	316
7		14	si	10	496
8		15	d	10	377
9		15	d	11	113
10		16	n	3	388
11		18	je	11	404
12	RVF	8	n	9	182
13		9	Ju	6	144
14		10	a	6	143
15		12	so	14	237
16		12	so	11	658
17		12	so	10	290
18		13	st	9	237
19		13	st	8	469
20		15	d	14	352
21		15	m	5	64
22		15	d	13	297
23		15	d	12	251
24		15	d	9	308
25		17	n	2	349
26		18	ju	10	272
Mean		13		9.58	313
Total					8135

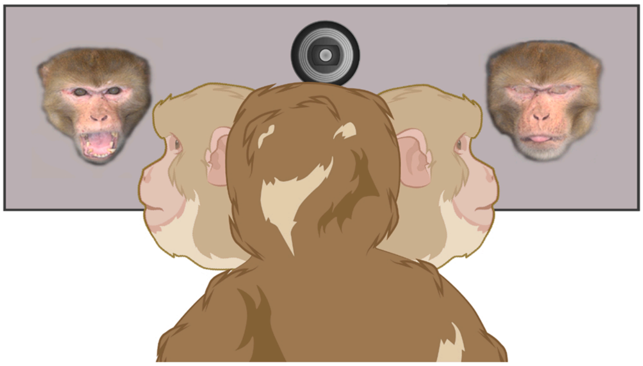

Fig. 1. Example of an attention bias preferential looking trial. Monkeys were shown a pair of social stimuli (one threat face and one neutral face, from a single unknown individual) presented on either side of a video camera. Head orientation was filmed.

time of writing, for non-human primates, no tool for detection of head orientation is available although the first tools for body posture (Bala et al., 2020; Labuguen et al., 2021), social interactions (Bala et al., 2020) and individual identity (Guo et al., 2020; Schofield et al., 2019; Witham, 2018) have recently been developed. Some of these models (e.g. Witham, 2018) provide facial landmark data, but none directly provide head pose estimation, specifically in real-time environment with noisy foreground.

More generally, platforms such as DeepLabCut Model Zoo (Mathis et al., 2018; Nath et al., 2019) offer a small but growing range of models for coding behavior in species ranging from horses to rodents, and including primates (e.g. Witham, 2018). As the number of models grows and users become familiar with these platforms computer vision scientific approaches, including DL models, will become standardised tools for behavioral coding in a range of disciplines. Model Zoo, for example, enables the extraction of detailed facial landmarks for the macaque face that can be stacked by a computer vision or deep learning model (as we present in the methods described here) for behavioral analysis in real time conditions.

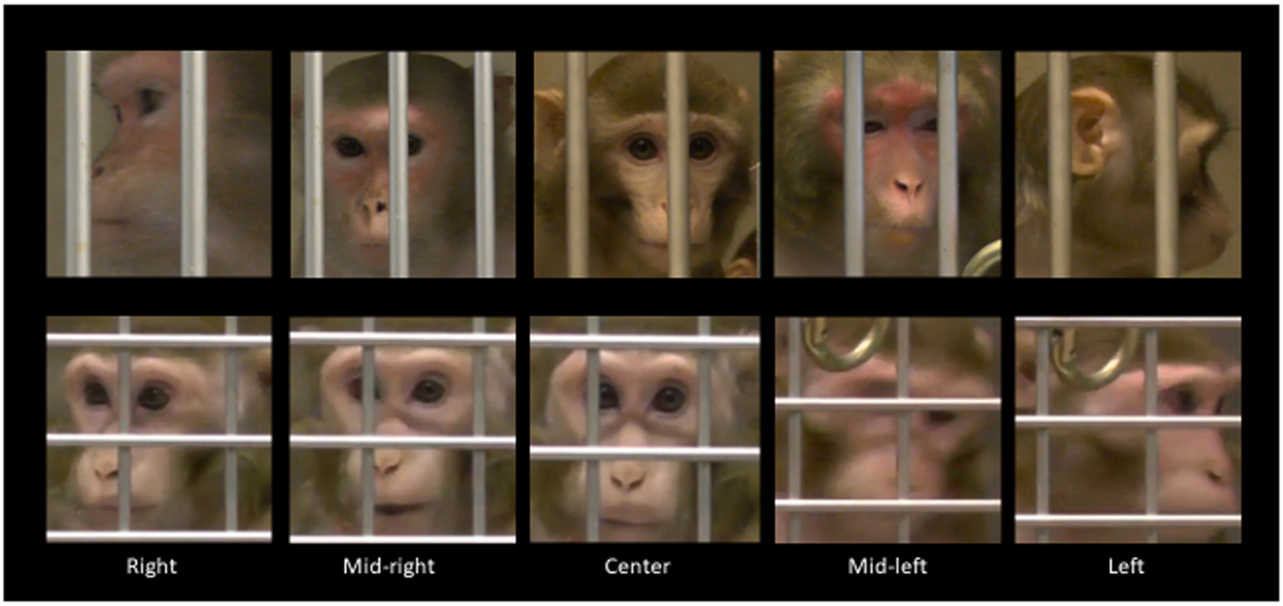

Fig. 2. Head orientation parameters for manual coding of frames using morphological features. ‘Center’ was coded when the face was oriented directly towards the camera, assessed by looking for equivalency in the size of the left/right nostrils (‘center’, lower panel). Where one or both nostrils were obscured by the bars we used left/right eye sockets and brow ridges, and visibility of the left/right ears as markers for orientation, allowing a small degree of error (‘center’, top panel shows maximum deviation permitted from absolute center). Where the head was turned so there was no longer equivalency in nostril size (or secondary markers where nostrils were not visible due to bars), orientation was considered not to be central. In this case, when the head was oriented so that both nostrils remained visible (assuming no occlusion by bars) and the furthest eye was not obscured by the nose ridge we applied a ‘mid’right’ or ‘mid-left’ code. When the head was turned so that only one nostril was visible (e.g. ‘left’, lower panel) and/or the bridge of the nose began to obscure the eye furthest from the camera (‘right’, lower panel) the frame was coded as ‘left’ or ‘right’ accordingly. Head orientation was coded in this way up to a maximal orientation of 90^0 relative to the camera (left, upper panel). Images selected to show individual differences in facial appearance, range of still frame quality (e.g. ‘mid-left’, lower panel shows blur caused by head movement) and visual obstruction by bars.

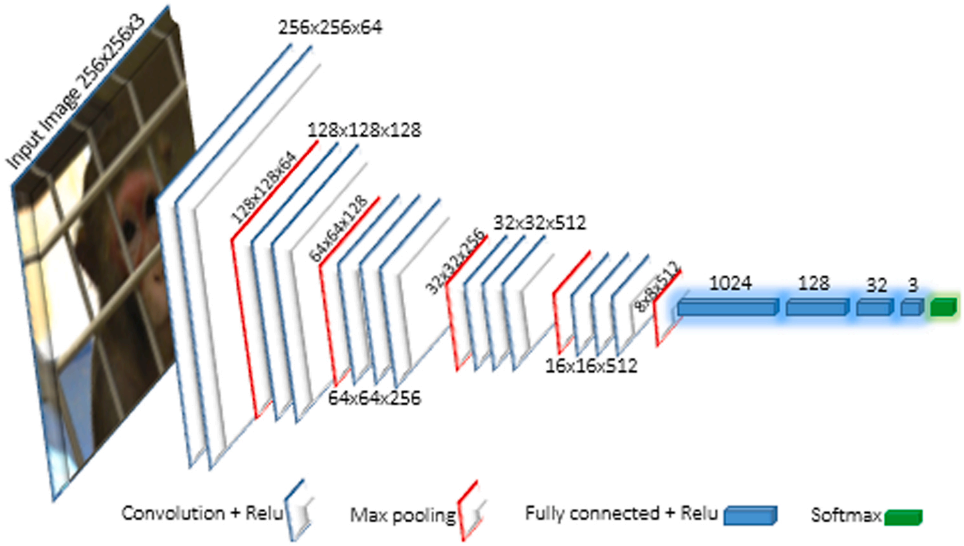

Fig. 3. An example of the DTL-HPE model applied to the VGG16 CNN network architecture (Simonyan and Zisserman, 2014) as the base model. The spatial pooling was constructed using 5 max-pooling layers.

Here we present, to our knowledge, the first open source model for head pose estimation in unrestrained rhesus macaques during behavioral psychological research. Importantly, our tool is robust to partial occlusion of the face, as occurs in real world unrestrained contexts (e.g. own and conspecific body parts, foliage in the wild and enclosure mesh in captivity).

2. Methods

2.1. Participants and video collection

Video of 26 adult female rhesus macaques *Macaca mulatta* housed in social breeding groups at the Centre for Macaques, MRC Harwell Institute, UK, was used in the current study (n = 26 animals, mean age = 13 years, range 8 – 18 years; group sizes 2–10: Table 1). Videos were of monkeys taking part in an attention bias preferential-looking task using a nonrestraint noninvasive approach, filmed with a Panasonic HC-V520 video camera (Fig. 1: Howarth et al., 2021). Monkeys had previously been trained, using positive reinforcement with food rewards, to station next to a target in the home enclosure to take part in cognitive testing (Kemp et al., 2017). Monkeys were always tested in their social group and were free to disengage and move away from the apparatus to join the rest of the group at any time during testing.

Each monkey originally took part in eight attention bias trials (mean = 7.8 trials, range = 4–8). Briefly, a pair of stimuli were loaded into an apparatus, with one image on either side of a video camera. When the monkey was oriented towards the camera, occluders covering the stimuli were removed and the monkey's gaze towards the two images was filmed. Video frame size was 640 × 480 pixels. For the current analysis we selected video for the first trial only, as our previous work shows most looking towards stimuli occurs on this trial for most monkeys (Howarth et al., 2021). A few monkeys have an avoidant attention profile, and since we selected videos for the first trial only, there was variability in the number of images available across individuals. We did not adjust for this as it was a natural characteristic of the data set. The 26 monkeys were selected at random from our existing database, full details of which are published in Howarth et al. (2021) where a detailed protocol and animal information for the full cohort can be accessed from the Supplementary material (Study 1).

2.2. Video annotation and dataset preparation

Each video was manually annotated on a frame-by-frame basis for orientation of the head with respect to the video camera. Boundaries for classes of head pose were based on visual assessment of the symmetry of morphological cues – primarily the nostrils and secondarily the eyes, ears, brow-ridge and nose-ridge (Fig. 2). Relative nostril size was used primarily as it was least likely to be obscured by the bars, was relatively robust to head elevation (which we did not measure) and was therefore the easiest measure to qualify for manually coding a large number of video frames of freely moving animals. We identified five classes for coding purposes, of which three classes were subsequently used for training and testing the DL model: ‘left’, ‘center’ and ‘right’. Our classes align with neurophysiological work showing distinct neuronal responses to different head orientations in macaques (Murphy and Leopold, 2019; Taubert et al., 2020). In those studies macaque neuronal responses to a digital macaque avatar in which the head orientation was experimentally and precisely manipulated for orientation (including elevation) were measured. There were distinct neuronal responses to faces oriented at 0^0 (i.e. directly oriented towards the viewer; Fig. 2, ‘center’) compared to when the stimulus head was oriented 30^0 to the left or right (equivalent here to our minimum threshold for ‘left’ and ‘right’; Fig. 2: ‘right’, lower panel). As in (Murphy and Leopold, 2019) we set $\pm 90^0$ as the upper boundary for head orientation (Fig. 2, ‘left’, upper panel). Because the monkeys in our study were unrestrained we allowed a small degree of error for ‘central’ which we assume to be $< \pm 10^0$. Faces that were visible but did not meet the criteria for ‘central’ or ‘left’/’right’ (i.e. approximately orientated at an angle $10^0 < 30^0$ relative to the camera), were classed as ‘mid-left’ and ‘mid-right’. Only frames in which the head and at least one eye were visible were annotated (except at the upper boundary for left/right $\pm 90^0$ where it was possible no eye was visible).

2.3. Deep transfer learning based head pose estimation (DTL-HPE)

In image processing tasks, extensive analysis indicates that deep learning (DL) provides an excellent evaluation method when used for large datasets (Soumare et al., 2021). Deep learning is a type of machine learning algorithm constructed from multi-layer neural networks that has many hidden layers, with a number of artificial neurons that can provide mathematical operations on the input dataset (Zou et al., 2019). There are various DL algorithms, among them CNN, which is considered the state-of-the-art approach to image classification at the time of writing. CNN simulates natural brain processing, as well as representing visual information among adjacent pixels and objects (Rawat and Wang, 2017).

Deep learning neural network architectures show improved versatility in performance when benchmarked against conventional ML approaches, because DL can work with unstructured datasets while ML is suited to structured data (Lei et al., 2020). ML models are trained on large, labelled datasets for which they show strong performance, but they show poorer performance when transferred to real-world applications because in real-world applications such labelled datasets are not available (Lei et al., 2020). By contrast, deep transfer learning (DTL) is an application of DL in which the knowledge gained through DL in one set of learning scenarios is transferred to others (Pan and Yang, 2009). The source domain represents the domain in which the knowledge is learned, while the target domain is the one to which the knowledge is transferred.

There are various DL architectures that have been developed for video and image processing, and the most advanced and accurate of these are those that utilise CNN. A number of open-access DL CNN architectures are available that have been trained on huge image datasets and that are suitable for use as base models in DTL. For the current study we selected four widely used CNN architectures as proposed base models, all of which have been evaluated for performance using the well know ImageNet Large-Scale Visual Recognition Challenge (ILSVRC: Berg et al., 2010) in which models compete in image recognition tasks from the ImageNet database of > 15million labelled high resolution images. The best performing and most widely utilized of these are: VGG-Nets VGG-16 and VGG19 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2016), and GoogLeNet’s Inception v3 (Szegedy et al., 2015). VGG-16 and VGG-19 are 16 layer and 19 layer CNN respectively. Both models have been well validated for transfer learning (Carvalho et al., 2017) and image classification (Mateen et al., 2019). ResNet50 is a 50 layer CNN which utilises the concept of skip connection allowing the feeding of the input data from previous layers to the next ones without modification. ResNet50 is also known as a residual network. The network utilises 1 × 1 convolutional layers, reducing the computational complexity by the elimination process. GoogLeNet is another CNN architecture that has two different versions namely Inception-v1 and Inception-v3, consisting of 42 layers (Alom et al., 2018).

The full DTL based model for HPE (from herein DTL-HPE) in rhesus macaques is available at https://osf.io/3npq8/. We trained the DTL-HPE model using each of the four base models over the recursive train/test partitions of the dataset using the code provided in Supplementary Material (Code S1). Specific parameters and experimental settings are also provided (Table S1). Fig. 3 shows the DTL-HPE model applied to the VGG16 CNN network architecture as the base model for transfer learning to occur. In this model the CNN is set to 256 × 256 RGB image which is forwarded to a stack of convolutional layers (in this case 16 layers).

Algorithm 1 shows the overall steps for the proposed DTL based HPE within the video data. Multiple baseline experiments were conducted by a number of classification trials to compare the HPE performances of VVG16, VVG19, ResNet50 and Inception-V3 based DL models. Initially, multiple train/test trials were run by partitioning the entire dataset randomly into training and testing proportions comprising 70% and 30% of the total images respectively. This analysis allowed us to assess transfer learning accuracy with the 26 monkeys used in training and testing. Using images from the same individuals in training and testing is commonly done but is likely to result in non-independent data (i.e. adjacent frames are non-independent). This makes a standard approach like cross-validation liable to produce biased classification outcomes, failing to indicate the utility of the model for transfer to new individuals and image sets.

Algorithm 1. Proposed methodology for head pose estimation (HPE).

```
Input: $V = \{ v \mid v$ is a video$\}$
Output: $hp \Rightarrow hp \in HP$ and $HP = \{Left, Right, Center\}$ is the head position set
Step 1: Data Preparation
Let $f$ represent the image frame
$\forall v \in V$, annotate $v$ and find $h \Rightarrow h \in HP$ using morphological features as:
    For each $f$ in a $v$:
        IF $f$ is a good frame containing full face:
            IF equivalency in the size of the left/right nostrils:
                - $hp$ for current $f$ is Center
            ELSE IF both nostrils remain visible & furthest eye is not obscured by the
            nose ridge
                - $hp$ for current $f$ is mid-left/mid-right
            ELSE IF the head is turned so that only one nostril is visible
                - $hp$ for current $f$ is left/right
        ELSE Ignore current $f$ and move to next
    End Loop
Let Training = $\{tr \in \}$ representing annotated image frames extracted from randomly
selected 70% of $v$
Let Testing = $\{ts \in v \Rightarrow ts \notin \text{Training}\}$ representing image frames extracted from 30% of
remaining $v$
Step 2: Hyper-Tuning
Let $ML_d$ = {VGG16, VGG19, ResNet50, Inception-v3} a set of Deep learning models
Let $q \in QM$ = {Precision, Recall, F1-score, Accuracy, Macro Avg F1 Score}
$\forall ml \in ML_d$ , select optimal parameters empirically for:
    - hidden layers
    - select Image Size
    - Select Batch size
For training run $r_1$ to $r_{10}$:
    Foreach $ml \in ML_d$
        - Split $tr$ into training ($trSet$: 80%) and validation ($vSet$: 20%)
        - Train and validate $ml$ over $trSet$ and $vSet$ respectively, for 40 epochs
        - Test $ml$ over unseen frames extracted from $ts$
        - Find $hp$, measure and store $q$ for the current $r$
    End Loop
    - Update the $tr$ and $ts$ using random selection of $v$ in Step 1
End Loop
```

Table 2
Performance metrics. **TP:** Correctly classified images that belong to that class; **TN**: Correct rejection of images that do not belong to that class; **FP**: Incorrect classification of images to a class they do not belong to; **FN**: Incorrect rejection of images from a class they belong to.

Performance Metric	Description
$\boldsymbol{Recall}_{(C)} = \frac{TP_C}{TP_C + FN_C}$.	The percentage of images that were classified to class *C*, compared to all images that should have been classified into *C*.
$\boldsymbol{Precision}_{(C)} = \frac{TP_C}{TP_C + FP_C}$.	The percentage of images correctly classified for class *C*.
$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$.	Overall accuracy of the model for all classes.
$\boldsymbol{F1\ Score}_{(C)} = \frac{2 * Precision(c) * Recall(c)}{Precision(c) + Recall(c)}$	Harmonic-mean of *precision* and *recall* indicating success rate of the model for class *C*.
$\boldsymbol{Macro\ Average\ (F1\ Score)} = \frac{\sum_{c=1}^{3} F1 \quad Score_{(C)}}{3}$.	Average of each class's F-1 score independent to sample size per class.

To assess the accuracy of the HPE model for transfer of DL to previously unseen individuals, a leave-k-out (LKO) strategy was subsequently utilised for the partitioning of training and test data, which is one of the commonly used strategies in ML model evaluation in similar scenarios (Khan et al., 2021; Little et al., 2017). For this analysis, the training set comprised all image frames extracted from videos of 18 monkeys who were selected at random (i.e. 70% of the total video data), and the testing set comprised all image frames of the remaining 8 monkeys (i.e. 30% of the entire video data) for the testing. We conducted 10 recursive runs using the random LKO strategy for the four DL base models to investigate the reliability and generalisability of the proposed DTL-HPE model. Statistical outcomes were generated following the parametric configurations detailed in Supplementary Material (Table S1) to classify head pose within the unseen image frames.

The final parametric configurations were set empirically based on several recursive trials over a random train/test partition (70%:30% respectively using LKO). In total there were 8135 image frames. The training set comprised 5114 images (961, 2622, 1531 images for training left, center and right classes respectively). The test-set contained 3021 images (343, 1571 and 1107 for testing left, center and right classes respectively). Because we selected videos at random, there were fewer training and test images with left orientation likely reflecting our finding of visual field effects in the primary study (Howarth et al., 2021). In that study monkeys showed reduced interest, or possible avoidance, of threat faces presented to the left visual field, and greater bias towards threat faces presented to the right visual field.

3. Results

3.1. Performance of the DTL-HPE model

Standard statistical metrics (i.e. recall, precision, accuracy, F1 score, macro average) were used to evaluate the classification performance based on confusion matrices retrieved from the proposed DTL based head pose classifiers (Table 2).

Table 3
Training, validation and testing performances of VVG16, VVG19, ResNet50 and Inception-V3 using 70% and 30% random training and testing partitions respectively. Results were equivalent for left, center and right HPE.

Model Name	Training Accuracy	Validation Accuracy	Test Accuracy
VVG16	1	0.99	1
VVG19	1	0.99	0.99
ResNet50	1	0.99	0.99
InceptionV3	1	0.98	0.98

3.2. Head pose estimation in known individuals

Table 3 shows the statistical outcomes from DTL-HPE obtained from the four base models, when trained and validated using 70% of the images and tested on 30% of the images. The DTL-HPE model achieved almost 100% accuracy in training, validation and testing with all four base models. Fig. 4 shows number of epochs to reach peak training and validation accuracy for each base model. While testing was performed on unseen images, the dependent nature of images from the same individual may cause relatively higher accuracies.

3.3. Head pose estimation in unknown individuals

Table 4 shows the statistical metrics retrieved from DTL-HPE for the four base models when tested on individuals ($n = 6$) whose images were not used during training and validation ($n = 18$ monkeys used in training and validation). The outcomes indicate similar performances (0.88–0.90 HPE accuracy) by all models except Inception-V3 which produced lower overall accuracy (0.77 HPE accuracy). The best performing base model overall was VGG16, which had the highest overall accuracy of 0.90. Generally, VGG16 also showed the greatest precision, recall and F1-scores across the three classes of head pose, indicating it was the least biased of the base models by number of training images for each class. For example, images of head orientation to the left were under-represented in the dataset. For this class, VGG16 achieved 0.78 recall, compared to 0.71, 0.74 and 0.66 from VGG19, ResNet50 and Inception V3 respectively. Likewise, the Macro average F1 score for VGG16 was also the highest (0.89) as compared to other models.

We subsequently assessed the impact of image size on DTL-HPE performance when using the VGG16 base model (Table 5). Generally, the model had greater accuracy for larger image frames. Accuracy increased from 0.78 to 0.89 when image size was increased from 64 × 64–256 × 256 respectively. There was no improvement in performance between images sized 256 × 256 and images sized 512 × 512. This may be an artefact of the original video frame size (640 ×480 pixels) so that changing it to 512 × 512 was not useful in this case. However, we might expect performance to improve with higher resolution images. In our work, we standardized all images to 256 × 256 pixels, to allow interpretation with respect to real time scenarios with low resolution images as well as consuming low computational resources.

Finally, to validate the performance of our DTL-HPE model for images of previously unseen individuals, we performed 10 recursive runs using the LKO strategy. Table 6 shows the statistical outcomes of the DTL-HPE model, using the VGG16 base model, for each of the 10 iterations. This indicates reliability and generalization of HPE model with a best accuracy of 93 %, with a fair precision (left 89 %, center 92 %, right 96 %), recall (left 94 %, center 96 %, right 88 %) and F1-score (left 92 %, center 94 %, right 92 %) for the three classes of head pose. The grand

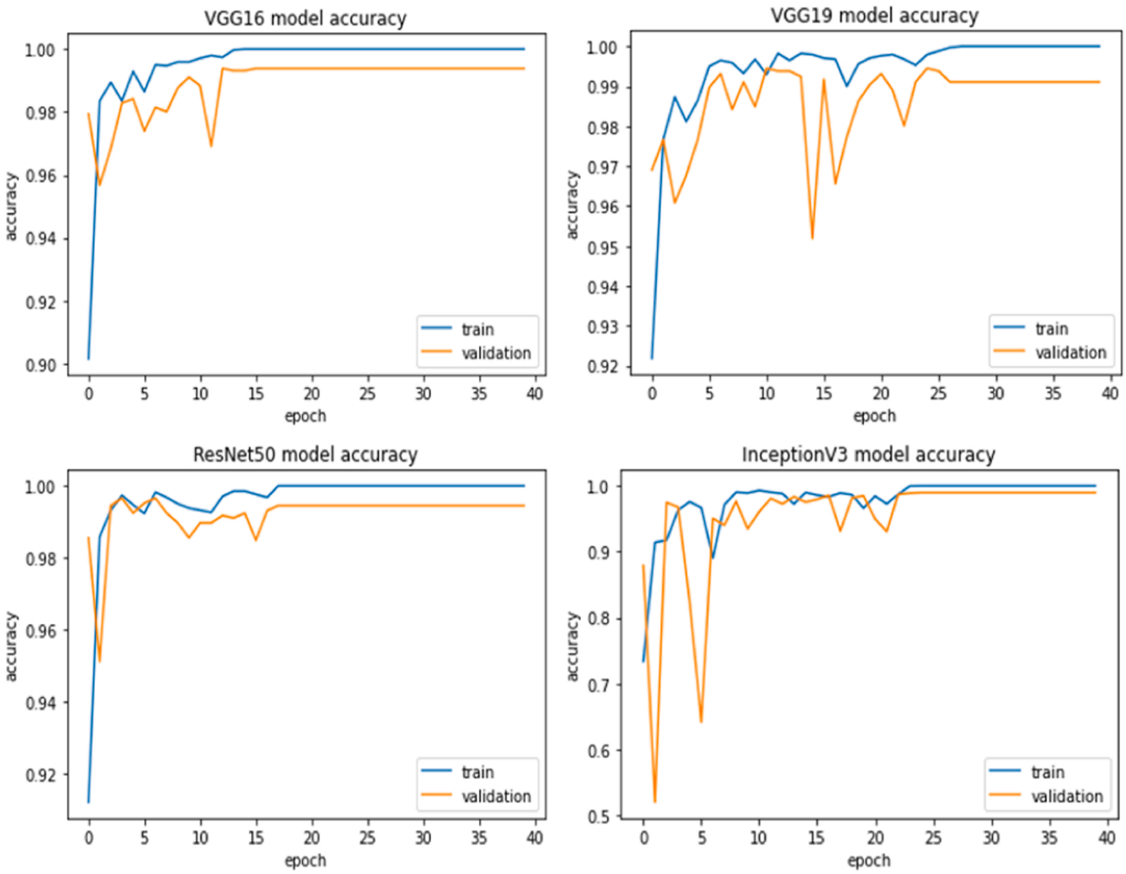

Fig. 4. Convergence of Proposed DTL based HPE classifiers over random partition of training and validation samples.

average accuracy of 10 recursive experiments was 89.5 % accuracy for overall three classes with 85 %, 91 % and 88 % of F1-score for left, center and right classes respectively. The model indicated slightly better performance for the center class, which may indicate a bias due to sample size, as center was over-represented in the dataset. Furthermore, varying number of samples per individual subject (i.e., video frames per monkey, see Table 1) may influence the model performance as a confounding factor.

4. Example application of the DTL-HPE model to assess hemispheric lateralization during a cognitive task

To illustrate the potential application of the DTL-HPE model for assessing hemispheric specialization in viewing preferences for social stimuli, we examined how the three classes of head pose assigned by the model mapped onto the location of the threat face (left/right visual field) that was shown during each video. Because of the small samples size (n = 26 monkeys) and because we included some video frame images from before and after the start of the attention bias trial to maximize the number of image frames for training and testing the model, we refer the reader to (Howarth et al., 2021) for a more complete analysis of the full data set (n = 108 monkeys), which was manually coded for eye gaze (but not head pose). Our example illustrates how the DTL-HPE model can be used currently, and its promise for future development for integration with eye pupil localization to estimate eye gaze.

Analysis was conducted in R (RCoreTeam, 2019). We constructed a generalised linear mixed effects model (GLMM) using R package ‘lme4’ version 1.1–15 (Bates et al., 2015). The response variable was number of

Table 4
Average outcomes using random LKO 10 iterations with VGG16, VGG19, ResNet50 and InceptionV3 for the DTL-HPE classes (left, center, right).

Model Name	Precision	Recall	F1-score	Accuracy	Macro Avg. F1-score
VGG16	L: 0.94 C:0.89 R:0.91	L:0.78 C:0.95 R:0.85	L:0.85 C:0.91 R:0.88	0.90	0.89
VGG19	L: 0.93 C:0.84 R:0.91	L:0.71 C:0.94 R:0.80	L:0.81 C:0.89 R:0.85	0.88	0.87
ResNet50	L: 0.98 C:0.86 R:0.91	L:0.74 C:0.96 R:0.83	L:0.84 C:0.90 R:0.87	0.89	0.87
InceptionV3	L: 0.50 C:0.88 R:0.73	L:0.66 C:0.79 R:0.77	L:0.57 C:0.83 R:0.75	0.77	0.72

Table 5
Impact of image size on DTL based HPE using VGG16 base model.

Image Size	Precision	Recall	F1-score	Accuracy	Macro Avg. F1-score
512 × 512	L: 0.94 C:0.92 R:0.84	L:0.77 C:0.91 R:0.90	L:0.84 C:0.91 R:0.87	0.89	0.88
256 × 256	L: 0.86 C:0.91 R:0.87	L:0.84 C:0.90 R:0.88	L:0.85 C:0.90 R:0.88	0.89	0.88
224 × 224	L: 0.83 C:0.79 R:0.97	L:0.74 C:0.97 R:0.67	L:0.78 C:0.87 R:0.79	0.84	0.82
128 × 128	L: 0.51 C:0.92 R:0.86	L:0.88 C:0.75 R:0.91	L:0.65 C:0.82 R:0.88	0.82	0.80
64 × 64	L: 0.65 C:0.79 R:0.81	L:0.65 C:0.85 R:0.72	L:0.65 C:0.82 R:0.76	0.78	0.74

frames, the predictor variables were entered as an interaction term: HPE (left, center, right) and visual field to which threat face was presented (left/right). Animal ID was entered as a random effect with a poisson family error distribution and log-link function specified. Model fit was assessed by visual inspection of plots of residuals. Model validity was assessed by comparing it against the null model (an identical model except for the removal of the predictor and control variables, with an intercept of 1 specified) using the anova command in R (Burnham and Anderson, 2002). We tested for collinearity using the vif command in the package 'car' finding no evidence (all vifs <1.54).

The full model including the interaction between visual field and HPE explained the data better than the null ($\chi^2 = 1921$, df $= 5$, $P < 0.001$). There was a significant 3 × 2 interaction between head pose and visual field (LRT=318.92, df=2, $P < 0.001$: Fig. 5), with significant 2 × 2 interactions between each combination of HPE (LxC, RxC and LxR) and visual field (all $z > 9.6$, all $P < <0.001$). There was a main effect of head pose (LRT=1600, df = 2, $P < 0.001$), again with significant 2 × 2 interactions between each combination of head pose (LxC, RxC and LxR: all $z > 5.74$, all $P < 0.001$). Visual inspection of Fig. 5 indicates that for videos containing trials with the threat face presented on the left, there were more frames with head pose to the right than either a) frames with head pose to the left or b) number of right head pose frames for videos containing trials with the threat face on the right. For videos containing trials in which the threat face was presented on the right, the difference between left and right head, while still significant, was greatly attenuated due to greater number of frames with left head pose, and fewer frames with right head pose.

Table 6
Recursive runs of the DTL-HPE model using the VGG16 base model and original (imbalanced) training dataset, with testing on randomly selected leave-k-out subjects.

Train/Test Runs	Precision	Recall	F1-score	Accuracy
1	L: 0.99 C:0.85 R:0.93	L:0.61 C:0.97 R:0.85	L:0.75 C:0.91 R:0.89	0.89
2	L: 0.97 C:0.88 R:0.94	L:0.75 C:0.97 R:0.87	L:0.85 C:0.93 R:0.90	0.91
3	L: 1.00 C:0.86 R:0.86	L:0.68 C:0.94 R:0.81	L:0.81 C:0.90 R:0.84	0.87
4	L: 0.97 C:0.84 R:0.95	L:0.75 C:0.99 R:0.77	L:0.85 C:0.91 R:0.85	0.88
5	L: 0.86 C:0.91 R:0.87	L:0.84 C:0.90 R:0.88	L:0.85 C:0.90 R:0.88	0.89
6	**L: 0.89** **C:0.92** **R:0.96**	**L:0.94** **C:0.96** **R:0.88**	**L:0.92** **C:0.94** **R:0.92**	**0.93**
7	L: 0.90 C:0.89 R:0.89	L:0.85 C:0.93 R:0.83	L:0.87 C:0.91 R:0.86	0.89
8	L: 0.97 C:0.88 R:0.94	L:0.75 C:0.97 R:0.87	L:0.85 C:0.93 R:0.90	0.91
9	L: 0.89 C:0.89 R:0.91	L:0.87 C:0.94 R:0.85	L:0.88 C:0.91 R:0.88	0.90
10	L: 0.99 C:0.90 R:0.84	L:0.80 C:0.90 R:0.87	L:0.89 C:0.90 R:0.86	0.88
Avg. (10 runs)	**L: 0.94** **C:0.89** **R:0.91**	**L:0.78** **C:0.95** **R:0.85**	**L:0.85** **C:0.91** **R:0.88**	**89.5**

5. Discussion

The development of non-restraint non-invasive technologies for assessing animal behavior is essential for improving scientific outcomes and animal welfare in neuroscientific and cognitive behavioral research in the laboratory, as well as accuracy in field settings. In the laboratory setting, new technologies show promise for replacing current methods that require restraint and invasive procedures, resulting in welfare benefits. Here, for the first time to our knowledge, we demonstrate the effective application of a deep learning based head pose estimation (DTL-HPE) model trained and tested in a non-restraint, noninvasive set up, with *Macaca mulatta* taking part in a cognitive task in their home enclosure. Monkeys were free to move away and rejoin the social group

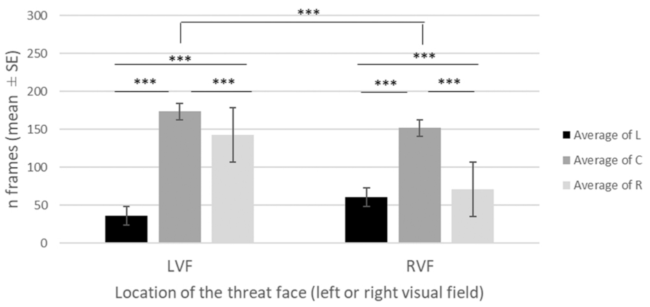

Fig. 5. Mean number of image frames (± SE) for which head pose was classed as left, center or right for n = 26 moneys. The x axis shows the location of the threat face within the apparatus (left or right visual field). Main interaction terms indicated for illustrative purposes only.

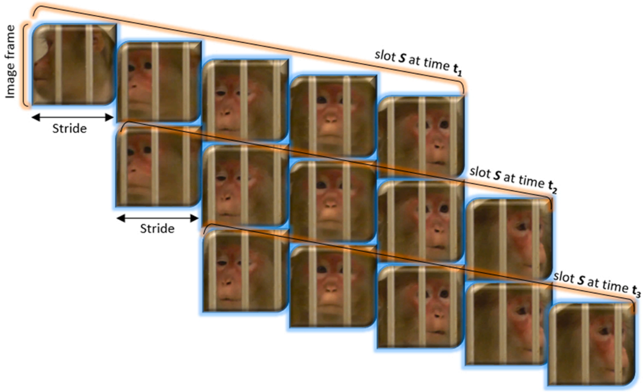

Fig. 6. Estimation of head pose over the fixed intervals of overlapping video slots. The head pose changes over time progression while the mode class calculated across slots represents the resulting head pose for the current slot. Shown progression from right to left head pose.

at any time, increasing the likelihood of blurring due to motion and obstruction of the face by components of the enclosure on the resulting video images, factors that historically have contributed to the rational for restraining animals. The model achieved up to 93% accuracy when tested with unknown individuals who were not used in the training set, with balanced outcomes for the three classes of head pose (92%, 94%, 92% accuracy for left, center and right orientation respectively). This finding demonstrates how DL can be applied as a non-invasive approach to assess head pose in unrestrained macaques taking part in cognitive tasks. Importantly, this approach is robust to the presumed noise created when animals are not restrained indicating potential 3Rs benefits.

We gained the greatest accuracy when we trained our DTL-HPE model using VGG-16 as the base model, although accuracy was only marginally better than VGG-19 and ResNet50. VGG-16-based CNN architectures have been successfully applied to detect non-human primate face cues in a few initial studies. For example, VGG-16 was used as a base model for identity and sex recognition in free-ranging chimpanzees, *Pan troglodytes verus*, attaining accuracies > 87% when images of individuals not used for training were used in the test set (Schofield et al., 2019). Charpentier et al. (2020) used VGG-16 based CNN architecture to identify individual Mandrill monkeys, *Mandrillus sphinx*, attaining accuracies > 83% for unseen individuals. Although the purpose of our model was different (to assess head pose, not individual identity) the accuracies achieved in these three studies support VGG-16 as one of a number of suitable base models for use in behavioral assessment using primate head and face cues. Additionally our DTL-HPE model retained accuracy at 89% for relatively low resolution image sizes of 256 × 256 pixels suggesting suitability for application where non-specialist filming equipment or remote filming of subjects may result in course-grain footage.

Our model extends the range of newly emerging models that apply DL, and machine learning (ML) more generally, to assess behavior in animals. Firstly, this is the first model to be published that is specifically designed to provide output about head pose for a nonhuman animal that we are aware of. It extends initial work by Mathis et al. (2018) who developed the DeepLabCut platform that provides landmark data for whole-body pose in a range of animal species. Here, we demonstrate an application of DL specifically tailored to head pose estimation in macaques engaged in cognitive tasks, an application with direct utility for contexts ranging from field playback experiments to neuroscientific studies of cognitive function. For such studies, models that provide direct output relating to HPE, avoid the further processing required of landmark data to interpret output in a meaningful way. With respect to early work in this field with macaques, Witham (2018) applied a ML algorithm for individual face recognition in rhesus macaques (54 facial landmarks), which allowed identification of individuals with > 85% accuracy. This is available to access on DeepLabCut (Mathis et al., 2018). More recently Labuguen et al. (2021) applied DL through the online platform DeepLabCut (Mathis et al., 2018) to detect facial landmarks in pictures of macaques gained from the internet and zoos (up to five facial landmarks: the eyes, ears and one landmark for the nose), providing a neural network for markerless whole-body pose estimation. In both cases, landmark coordinates are produced, and our model differs in the output specifically addressing head pose relative to the viewer (left, central, right), as would be useful, for example, in cognitive testing scenarios. Regarding obstruction, in Labuguen et al. (2021) it is likely that pictures from the internet were already pre-selected for good visibility of the face and, where there was obstruction of facial features the authors report that the approximate location of the obstructed facial feature was manually labelled, although no data are provided on the impact of obstruction on accuracy. Witham (2018), tested their face recognition algorithm with a subset of manually selected faces that were partially obscured or showed rotation of the head. These two factors led to a reduction in accuracy from > 90% in the larger image test set to 76% accuracy for obscured faces, and to 60% accuracy for rotated heads. There are clear synergies between the approaches in terms of application to refine quality and quantity of information gained in output, especially where head pose and obstruction have significant impact on other applications of DL such as individual recognition (see also Shukla et al., 2019; Sinha et al., 2019). DL models have also been applied to identify primate species other than macaques, including chimpanzees (Freytag et al., 2016; Schofield et al., 2019) and mandrills (Charpentier et al., 2020). These studies again focused on individual recognition and application for monitoring of habitat use and social systems. Hence, no DTL model for HPE in primates taking part in cognitive research has yet been developed and made available.

The application of a DTL-HPE model is potentially far-reaching in cognitive and psychological research with animals including humans. Head orientation has been used widely and reliably to assess direction of visual attention in primates in a number of psychological paradigms, e.g. gaze following (Ferrari et al., 2000), hemispheric lateralization of opponent viewing during agonistic interactions in the field (Casperd and Dunbar, 1996), hemispheric lateralization in acoustic processing (Teufel et al., 2007), kin recognition in acoustic field playback experiments (Pfefferle et al., 2014). Head turn is a commonly used measure of lateralization in a number of domesticated animals (Siniscalchi et al., 2021), and in species with laterally placed eyes who typically turn the head to view objects of interest (Rogers, 2010; Siniscalchi, 2021). In primates the literature indicates a general left hemisphere (right visual field) bias in approach and exploratory behavior, and a right hemisphere (left visual field) bias in avoidance of threatening stimuli (Vallortigara and Rogers, 2005) which can provide information about internal affective states that in turn could be used to both identify and improve welfare in vulnerable animals (Rogers, 2010). In our illustrative application of our model, we found an overall bias in head pose to monkeys' right hand sides, suggesting a right hemisphere (left visual field) priority of processing conspecific threat-neutral face pairs. This finding fits with the general pattern reported in the literature for hemispheric specialization in information processing of socially relevant stimuli including faces (Rogers, 2010), and specifically the role of the right hemisphere in avoidance of threat (Vallortigara and Rogers, 2005). The exemplar data presented here reflect the pattern seen in our larger dataset based on manual coding of eye gaze direction from 108 rhesus macaques revealing an avoidant attentional bias away from threatening stimuli presented to the left visual field, which is particularly enhanced during initial trials (Howarth et al., 2021).

6. Future directions

We identify several future directions for DTL models to assess behavioral indicators of attention such as head pose in nonhuman animals. One key future development is to refine the model for micro level behavior analysis such as eye movements and gaze direction. Social diurnal primates including humans have dedicated and separate neural circuitries for processing visual gaze cues including head orientation (Deaner and Platt, 2003; Hadjidimitrakis, 2020; Taubert et al., 2020; Wilson et al., 2000) and eye gaze (Deaner and Platt, 2003; Langton et al., 2000; Sparks, 2002). Each cue therefore has its own informational value, although their signal value is intrinsically linked (Ferrari et al., 2000; Hadjidimitrakis, 2020; Itti et al., 2003). Head restraint in primate cognitive research may be combined with other invasive methods such as surgically implanted scleral coils (Judge et al., 1980) to record eye gaze to visual stimuli e.g. (Adade and Das, 2019; Arora et al., 2019). A current limitation in the development of non-invasive eye-tracking devices is the lack of tools to integrate information on head orientation with information on pupil location, to triangulate direction of eye-gaze (Hopper et al., 2020). Existing image processing methods rely on facial landmark detection which requires an unobstructed view of most, or all, facial landmarks for reliable triangulation of coordinates (Khan et al., 2020). Reliability is therefore particularly impacted by blurred or obscured areas of the face (e.g. Fornalczyk and Wojciechowski, 2017; María Díaz Barros et al., 2017). Here we present a response to this challenge with a model for HPE on which further refinements to assess pupil location, and subsequently eye gaze direction, can be built, both in human and animal studies.

An additional development to explore is the use of slot-level analysis (Khan et al., 2021) to deal with boundaries between the three classes of head pose we utilized in the current study (Fig. 6). We limited our model to three classes of head pose due to the small data set available for the intermediate head positions (i.e. mid-left and mid-right). In nature, the transition from one pose to other is a progression over time; transition from left to right head pose occurs over a certain time interval and consequently includes multiple image frames within that ‘slot’. The number of images depends on speed of movement and frame rate. Utilization of slot level analysis rather than instance (i.e. single frame) based analysis in real time and video streamed data would allow for intermediate states between poses to be accounted for, as has been successfully applied by Khan et al. (2021).

Time slot analysis works as follows. Instead of specifying intermediate classes such as mid-left and mid-right, trained on individual images (and irrespective of what the preceding and following images are labelled as), the slot based video analysis considers images in their sequence. Fig. 6 demonstrate the example of slots containing five image frames indicating the transitions from right to left head pose. A DL based HPE model that utilizes time slot analysis would classify each image frame within a slot, resulting in a mode value for HPE outcomes (i.e. the head pose with the maximum counts within and across consecutive slots). This may also overcome the misclassifications by single-image-frame models that can be caused by real time dynamics such as a blurred images caused by movement, or acute occlusion. This would generate more reliable progressive HPE outcomes. This approach might be useful specifically for live and pre-recorded video data analysis in unrestrained animals such as those we worked with who were free to move around throughout testing.

In conclusion, we hope that the DTL approach presented here will provide direction for the future development of remote gaze estimation in animals, and alternative approaches to current head restraint practices in laboratory contexts. With good animal husbandry practices in captivity, including station training animals to engage with research protocols while housed with and retaining access to the social group, DTL approaches should allow researchers, funding bodies, ethics committees and other review panels to identify situations in which head restraint may no longer be required.

Funding

Collection of video material was financially supported by the National Centre for the 3Rs grant NC/L000539/1 to E.J. Bethell.

Ethics approval

Protocols for collection of the video material were developed following discussion with the facility Home Office Inspector (Nov 2011) and carried out in accordance with ethical guidelines for work with nonhuman primates (NC3Rs, 2006, 2015). Approval for this Study was granted by the Medical Research Council Animal Welfare and Ethical Review Body (AWERB) in 2014, Roehampton University Ethics Committee (approval #LSC 14/113) and the LJMU ethics panel (approval #EB/2014–1). Animal health was monitored daily by the care staff at CFM, and annually with a full veterinary examination. Methods and results are reported according to the ARRIVE guidelines (Kilkenny et al., 2010).

Code availability

Full model code is available to download at https://osf.io/3npq8/.

Open Access

The data and materials for all experiments are available in Supplementary Material, all of which can be downloaded at https://osf.io/3npq8/. Videos are not published due to data protection but may be available from EJB upon reasonable request.

Data Availability

All supporting materials, including the DL trained model, parameter configurations (Table S1), and model code (Code S1) are available at https://osf.io/3npq8/. Cognitive data are available from EJB on reasonable request.

Acknowledgements

We thank Caralyn Kemp for collection of video material, and Harriet Thatcher and Isabelle Szott who assisted. We also thank the staff at CFM, particularly Claire Witham and David Farningham for organizational support, and Faye Peters and Sebastian Merritt who additionally assisted with animal training and data collection.

Conflicts of interest/Competing interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Authors’ contributions

For collection of video material, EJB conceived the study, supervised the original project and preprocessed video. WK and EJB manually coded the videos for head orientation. WK and AH conducted data processing, analysis and implementation of the DTL-HPE model. All authors wrote their respective background and methods sections and all contributed to the final draft of the manuscript.

Declarations

Consent to participate

N/A.

Consent for publication

All authors consent to publication.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2022.105708.

References

Adade, S., Das, V.E., 2019. Vertical vergence in nonhuman primates depends on horizontal gaze position. Strabismus. Taylor & Francis Online, pp. 172–181. https://doi.org/10.1080/09273972.2019.1629465.

Adams, D.L., Economides, J.R., Jocson, C.M., Horton, J.C., 2007. A biocompatible titanium headpost for stabilizing behaving monkeys. J. Neurophysiol. 98 (2), 993–1001. https://doi.org/10.1152/jn.00102.2007.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B. C., Awwal, A.A.S., Asari, V.K., 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv Prepr 1803, *01164*.

Arora, H.K., Bharmauria, V., Yan, X.G., Sun, S.H., Wang, H.Y., Crawford, J.D., 2019. Eye-head-hand coordination during visually guided reaches in head-unrestrained macaques. J. Neurophysiol. 122 (5), 1946–1961. https://doi.org/10.1152/jn.00072.2019.

Bailly, K., Milgram, M., 2009. Boosting feature selection for neural network based regression. Neural Netw. 22 (5–6), 748–756.

Bala, P.C., Eisenreich, B.R., Yoo, S.B.M., Hayden, B.Y., Park, H.S., Zimmermann, J., 2020. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio, 12, Article 4560 *Nat. Commun.* 11 (1). https://doi.org/10.1038/s41467-020-18441-5.

Bates, D., Machler, M., Bolker, B.M., & Walker, S.C. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48. Go to ISI://WOS:000365981400001.

Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Lin, J.C.-W., Fortino, G., 2021. Deep learning for pedestrian collective behavior analysis in smart cities: a model of group trajectory outlier detection. Inf. Fusion 65, 13–20.

Berg, A. , Deng, J. , & Fei Fei, L. (2010). Large scale visual recognition challenge (ILSVRC). https://image-net.org/challenges/LSVRC/2010/.

Berger, M., Agha, N.S., Gail, A., 2020. Wireless recording from unrestrained monkeys reveals motor goal encoding beyond immediate reach in frontoparietal cortex (Article). *Elife* 9 (29), e51322. https://doi.org/10.7554/eLife.51322.

Burnham, K.P., Anderson, D.R., 2002. A practical information-theoretic approach. Model Sel. Multimodel Inference *2*.

Carvalho, T., De Rezende, E.R., Alves, M.T., Balieiro, F.K., & Sovat, R.B. (2017). Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA),

Casperd, J.M., Dunbar, R.I.M., 1996. Asymmetries in the visual processing of emotional cues during agonistic interactions by gelada baboons. Behav. Process. 37 (1), 57–65. https://doi.org/10.1016/0376-6357(95)00075-5.

Charpentier, M.J.E., Harté, M., Poirotte, C., de Bellefon, J.M., Laubi, B., Kappeler, P.M., Renoult, J.P., 2020. Same father, same face: Deep learning reveals selection for signaling kinship in a wild primate. Sci. Adv. 6 (22), eaba3274 https://doi.org/10.1126/sciadv.aba3274.

Deaner, R.O., Platt, M.L., 2003. Reflexive social attention in monkeys and humans. Curr. Biol. 13 (18), 1609–1613. https://doi.org/10.1016/j.cub.2003.08.025.

Ferrari, P.F., Kohler, E., Fogassi, L., & Gallese, V. (2000). The ability to follow eye gaze and its emergence during development in macaque monkeys. Proceedings of the National Academy of Sciences, 97(25), 13997–14002. https://doi.org/10.1073/pnas.250241197.

Fornalczyk, K., & Wojciechowski, A. (2017, 3–6 Sept. 2017). Robust face model based approach to head pose estimation. 2017 Federated Conference on Computer Science and Information Systems (FedCSIS),

Freytag, A., Rodner, E., Simon, M., Loos, A., Kuhl, H.S., Denzler, J., 2016. Chimpanzee faces in the wild: log-euclidean CNNs for predicting identities and attributes of primates. In: Rosenhahn, B., Andres, B. (Eds.), Pattern Recognition, Gcpr 2016, Vol. 9796. Springer International Publishing Ag, pp. 51–63. https://doi.org/10.1007/978-3-319-45886-1_5.

Ghazanfar, A.A., Santos, L.R., 2004. Primate brains in the wild: the sensory bases for social interactions. Nat. Rev. Neurosci. 603–616. https://doi.org/10.1038/nrn1473.

Guo, S.T., Xu, P.F., Miao, Q.G., Shao, G.F., Chapman, C.A., Chen, X.J., He, G., Fang, D.Y., Zhang, H., Sun, Y.W., Shi, Z.H., Li, B.G., 2020. Automatic identification of individual primates with deep learning techniques. Article 101412 *iScience* 23 (8), 32. https://doi.org/10.1016/j.isci.2020.101412.

Hadjidimitrakis, K., 2020. Coupling of head and hand movements during eye-head-hand coordination: there is more to reaching than meets eye. *J. Neurophysiol.* 123 (5), 1579–1582. https://doi.org/10.1152/jn.00099.2020.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Hopper, L.M., Gulli, R.A., Howard, L.H., Kano, F., Krupenye, C., Ryan, A.M., Paukner, A., 2020. The application of noninvasive, restraint-free eye-tracking methods for use with nonhuman primates. Behav. Res. Methods. https://doi.org/10.3758/s13428-020-01465-6.

Howarth, E.R., Kemp, C., Thatcher, H.R., Szott, I.D., Farningham, D., Witham, C.L., Holmes, A., Semple, S., Bethell, E.J., 2021. Developing and validating attention bias tools for assessing trait and state affect in animals: a worked example with Macaca mulatta. Appl. Anim. Behav. Sci. 234, 105198.

Itti, L., Dhavale, D., Pighin, F., 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention 2003 doi: 10.1117/12.512618.

Judge, S.J., Richmond, B.J., Chu, F.C., 1980. Implantation of magnetic search coils for measurement of eye position: an improved method. Vis. Res. 20 (6), 535–538. https://doi.org/10.1016/0042-6989(80)90128-5.

Khan, W., Crockett, K., O'Shea, J., Hussain, A., Khan, B.M., 2021. Deception in the eyes of deceiver: a computer vision and machine learning based automated deception detection. Expert Syst. Appl. 169, 114341.

Khan, W., Hussain, A., Kuru, K., Al-Askar, H., 2020. Pupil localisation and eye centre estimation using machine learning and computer vision. Sensors 20 (13), 3785.

Labuguen, R., Matsumoto, J., Negrete, S.B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K., Shibata, T., 2021. MacaquePose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture (Article). Front. Behav. Neurosci. 14 (8), 581154. https://doi.org/10.3389/fnbeh.2020.581154.

Langton, S.R., Watt, R.J., Bruce, V., 2000. Do the eyes have it? Cues to the direction of social attention. Trends Cogn. Sci. 4 (2), 50–59.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K., 2020. Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech. Syst. Signal Process. 138, 106587.

Li, J., Wang, J., Ullah, F., 2020. An end-to-end task-simplified and anchor-guided deep learning framework for image-based head pose estimation. IEEE Access 8, 42458–42468.

Little, M.A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P., 2017. Using and understanding cross-validation strategies. Perspectives on Saeb et al. GigaScience 6 (5). https://doi.org/10.1093/gigascience/gix020.

Mandalaywala, T.M., Parker, K.J., Maestripieri, D., 2014. Early experience affects the strength of vigilance for threat in rhesus monkey infants. *Psychol. Sci.* 25 (10), 1893–1902. https://doi.org/10.1177/0956797614544175.

María Díaz Barros, J., Garcia, F., Mirbach, B., & Stricker, D. (2017, 17–20 Sept. 2017). Real-time monocular 6-DOF head pose estimation from salient 2D points. 2017 IEEE International Conference on Image Processing (ICIP),

Mateen, M., Wen, J., Song, S., Huang, Z., 2019. Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry 11 (1), 1.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21 (9), 1281. https://doi.org/10.1038/s41593-018-0209-y.

McCay, K.D., Ho, E.S., Shum, H.P., Fehringer, G., Marcroft, C., Embleton, N.D., 2020. Abnormal infant movements classification with deep learning on pose-based features. IEEE Access 8, 51582–51592.

Murphy, A.P., Leopold, D.A., 2019. A parameterized digital 3D model of the Rhesus macaque face for investigating the visual processing of social cues. J. Neurosci. Methods 324, 108309. https://doi.org/10.1016/j.jneumeth.2019.06.001.

Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W., 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat. Protoc. 14 (7), 2152–2176.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Pfefferle, D., Ruiz-Lambides, A.V., & Widdig, A. (2014). Female rhesus macaques discriminate unfamiliar paternal sisters in playback experiments: support for acoustic phenotype matching [Article]. Proceedings of the Royal Society B-Biological Sciences, 281(1774), 8, Article 20131628. https://doi.org/10.1098/rspb.2013.1628.

Prescott, M.J., Lidster, K., 2017. Improving quality of science through better animal welfare: the NC3Rs strategy. Lab Anim. 46 (4), 152.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. 29 (9), 2352–2449.

RCoreTeam. (2019). R: A language and environment for statistical computing. In R Foundation for Statistical Computing. https://www.R-project.org/.

Rogers, L.J., 2010. Relevance of brain and behavioural lateralization to animal welfare. Appl. Anim. Behav. Sci. 127 (1), 1–11. https://doi.org/10.1016/j.applanim.2010.06.008.

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., Carvalho, S., 2019. Chimpanzee face recognition from videos in the wild using deep learning. Sci. Adv. 5 (9), eaaw0736 https://doi.org/10.1126/sciadv.aaw0736.

Shukla, A., Cheema, G.S., Anand, S., Qureshi, Q., Jhala, Y., 2019. Primate face identification in the wild. In: Nayak, A.C., Sharma, A. (Eds.), Pricai 2019: Trends in Artificial Intelligence, Pt Iii, 11672. Springer International Publishing Ag, pp. 387–401. https://doi.org/10.1007/978-3-030-29894-4_32.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition arXiv Prepr. arXiv 1409 2014 *1556*.

Sinha, S., Agarwal, M., Vatsa, M., Singh, R., Anand, S., 2019. Exploring bias in primate face detection and recognition. In: LealTaixe, L., Roth, S. (Eds.), Computer Vision, 11129. Springer International Publishing Ag, pp. 541–555. https://doi.org/10.1007/978-3-030-11009-3_33.

Siniscalchi, M., d'Ingeo, S., Quaranta, A., 2021. Lateralized emotional functioning in domestic animals (Article). *Appl. Anim. Behav. Sci.* 237 (12), 105282. https://doi.org/10.1016/j.applanim.2021.105282.

Soumare, H., Benkahla, A., Gmati, N., 2021. Deep learning regularization techniques to genomics data. Array, 100068.

Sparks, D.L., 2002. The brainstem control of saccadic eye movements. Nat. Rev. Neurosci. 3 (12), 952–964.

Szegedy, C., Liu, W., Jia, Y. Sermanet, J., Reed, S. Anguelov, D., Erhan, D. Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9.

Taubert, J., Japee, S., Murphy, A.P., Tardiff, C.T., Koele, E.A., Kumar, S., Leopold, D.A., Ungerleider, L.G., 2020. Parallel processing of facial expression and head orientation in the macaque brain. J. Neurosci. 40 (42), 8119–8131. https://doi.org/10.1523/jneurosci.0524-20.2020.

Teufel, C., Hammerschmidt, K., Fischer, J., 2007. Lack of orienting asymmetries in Barbary macaques: implications for studies of lateralized auditory processing. *Anim. Behav.* 73, 249–255. https://doi.org/10.1016/j.anbehav.2006.04.011.

Valletta, J.J., Torney, C., Kings, M., Thornton, A., Madden, J., 2017. Applications of machine learning in animal behaviour studies. Anim. Behav. 124, 203–220. https://doi.org/10.1016/j.anbehav.2016.12.005.

Vallortigara, G., Rogers, L.J., 2005. Survival with an asymmetrical brain: advantages and disadvantages of cerebral lateralization. Behav. Brain Sci. 28 (4), 575–589. https://doi.org/10.1017/S0140525X05000105.

Wang, C., Song, X., 2014. Robust head pose estimation via supervised manifold learning. Neural Netw. 53, 15–25.

Wilson, H.R., Wilkinson, F., Lin, L.M., Castillo, M., 2000. Perception of head orientation. Vis. Res. 40 (5), 459–472. https://doi.org/10.1016/s0042-6989(99)00195-9.

Wilson, V.A.D., Kade, C., Moeller, S., Treue, S., Kagan, I., Fischer, J., 2020. Macaque gaze responses to the primatar: a virtual macaque head for social cognition research. *Front. Psychol.* 11 (1645) https://doi.org/10.3389/fpsyg.2020.01645.

Winters, S., Dubuc, C., Higham, J.P., 2015. Perspectives: the looking time experimental paradigm in studies of animal visual perception and cognition. Ethology 121 (7), 625–640.

Witham, C.L., 2018. Automated face recognition of rhesus macaques. J. Neurosci. Methods 300, 157–165. https://doi.org/10.1016/j.jneumeth.2017.07.020.

Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2017). Towards large-pose face frontalization in the wild. Proceedings of the IEEE international conference on computer vision,

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A., 2019. A primer on deep learning in genomics. Nat. Genet. 51 (1), 12–18.