iScience, Volume *23*

Supplemental Information

Automatic Identification of Individual Primates with Deep Learning Techniques

Songtao Guo, Pengfei Xu, Qiguang Miao, Guofan Shao, Colin A. Chapman, Xiaojiang Chen, Gang He, Dingyi Fang, He Zhang, Yewen Sun, Zhihui Shi, and Baoguo Li

Supplementary Information - Trasparent Methods

Table S2. The ablation experiments for Tri-attention network, related to Figure 1 and 2. The individual identification accuracies obtained by Object level attention model and Partial level attention model and Tri-attention network for 17 primate species with more than 19 individuals and four non-primate species.

Animal species	The channel of **Object level attention model**	The channel of **Partial level attention model**	Tri-attention network
Weeper Capuchin (*Cebus olivaceus*)	0.9010	0.7802	0.9305
Black-capped Capuchin (*Cebus apella*)	0.8904	0.9045	0.9627
Rhesus Macaque (*Macaca mulatta*)	0.9101	0.9322	0.9126
Common Chimpanzee (*Pan troglodytes*)	0.8357	0.8231	0.9358
Common Squirrel Monkey (*Saimiri sciureus*)	0.8928	0.8968	0.9083
Green Monkey (*Chlorocebus sabaeus*)	0.8449	0.8992	0.9477
Mandrill (*Mandrillus sphinx*)	0.8000	0.9182	0.9294
Golden Snub-nosed Monkey (*Rhinopithecus roxellana*)	0.8845	0.9340	0.9412
François' langur (*Trachypithecus francoisi*)	0.8571	0.8571	0.9014
Olive Baboon (*Papio anubis*)	0.9264	0.7341	0.9341
Ring-tailed Lemur (*Lemur catta*)	0.9193	0.7897	0.9457
De Brazza’s Monkey (*Cercopithecus neglectus*)	0.9326	0.9299	0.9508
Patas Monkey (*Erythrocebus patas*)	0.9255	0.9184	0.9562
Greater Spot-nosed Monkey (*Cercopithecus nictitans*)	0.9014	0.8309	0.9543
Northern White-cheeked Gibbon (*Nomascus leucogenys*)	0.9132	0.9077	0.9342
Tibetan Macaque (*Macaca thibetana*)	0.9430	0.8134	0.9504

Crab-eating Macaque (*Macaca fascicularis*)	0.8965	0.7471	0.9337
Meerkat (*Suricata suricatta*)	0.9010	0.7796	0.9013
Lion (*Panthera leo*)	0.9192	0.6563	0.9355
Red Panda (*Ailurus fulgens*)	0.9170	0.8264	0.9216
Tiger (*Panthera tigris*)	0.9256	0.8807	0.9438

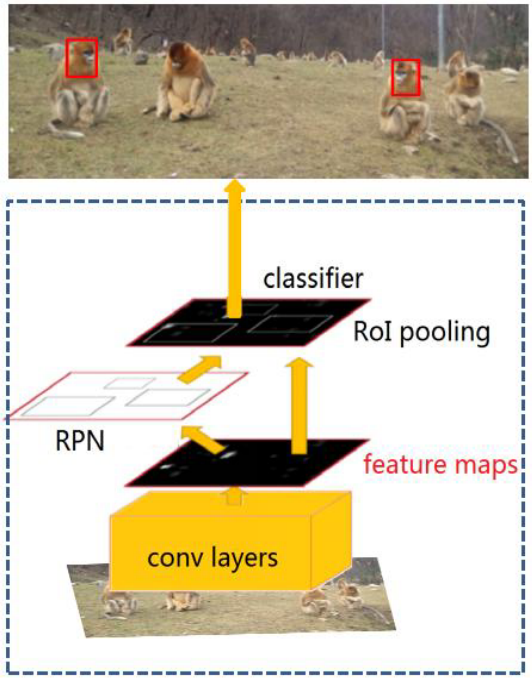

Figure S1. The face detection of golden snub-nosed monkeys by Faster RCNN, related to Figures 2, 4 and 5. We can detect the monkeys' faces when they show their faces. However, the other two individuals lower their heads in this moment, which results in Faster RCNN cannot find their faces.

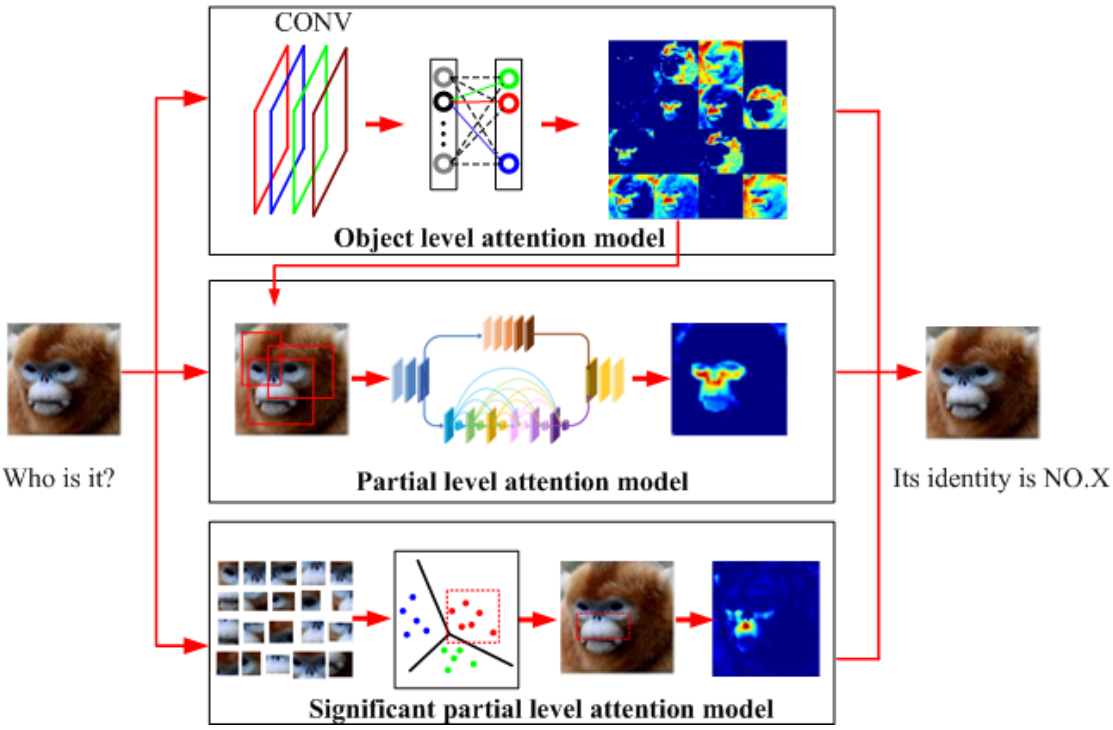

Figure S2. The framework of Tri-attention network, related to Figure 2. Object level attention model is used mainly to extract the features of the whole facial image. Significant partial level attention model pays attention to capture the local feature of a relatively fixed facial skin area. Partial level attention model focuses on the specific feature of a restricted smaller area, and different individuals would have their own specific facial areas.

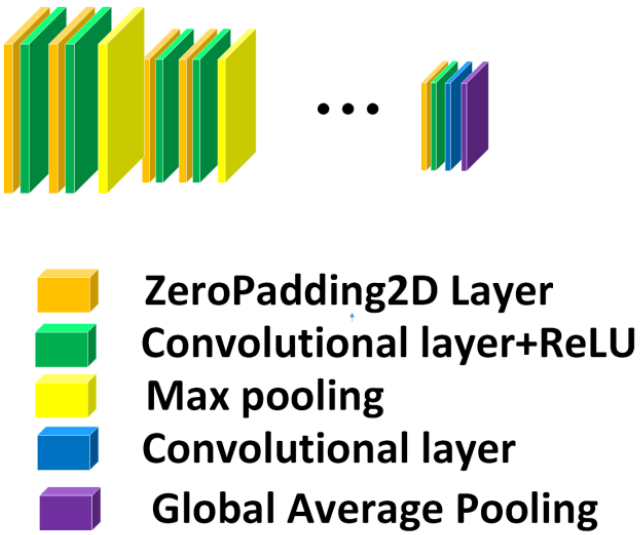

Figure S3. Object level attention model, related to Figure 2.

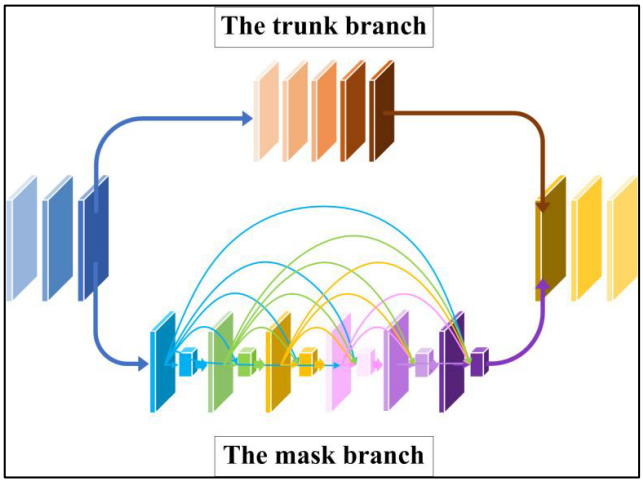

Figure S4. The structure of residual network with attention mechanism, related to Figure 2.

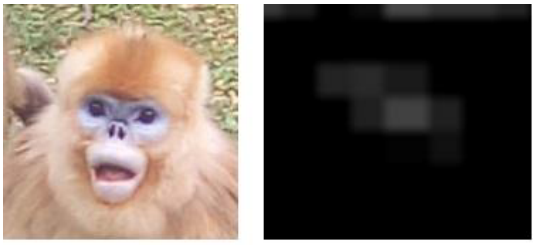

Figure S5. The feature map shows the model pays attention to feature extraction from the "skin area", related to Figure 2. (a) The facial image of golden snub-nosed monkey. (b)The feature map for the attentional facial region of the "skin area".

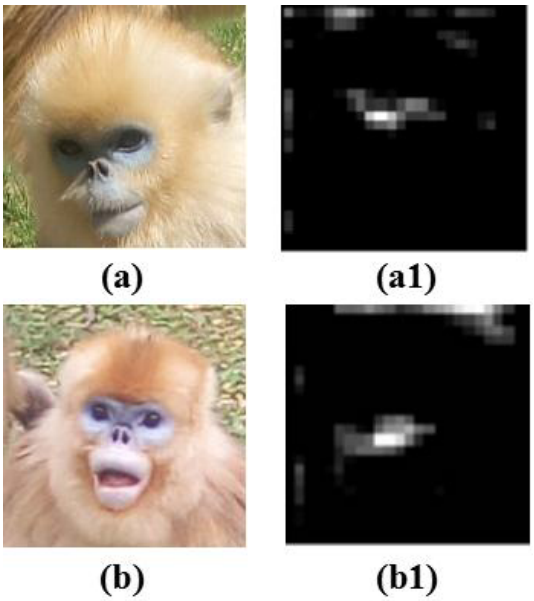

Figure S6. The original facial images of golden snub-nosed monkeys and the feature maps extracted by significant partial level attention model, related to Figure 2. (a) The facial image of golden snub-nosed monkey. (a1) The feature map for the attentional facial region. (b) The facial image of golden snub-nosed monkey. (b1) The feature map for the attentional facial region.

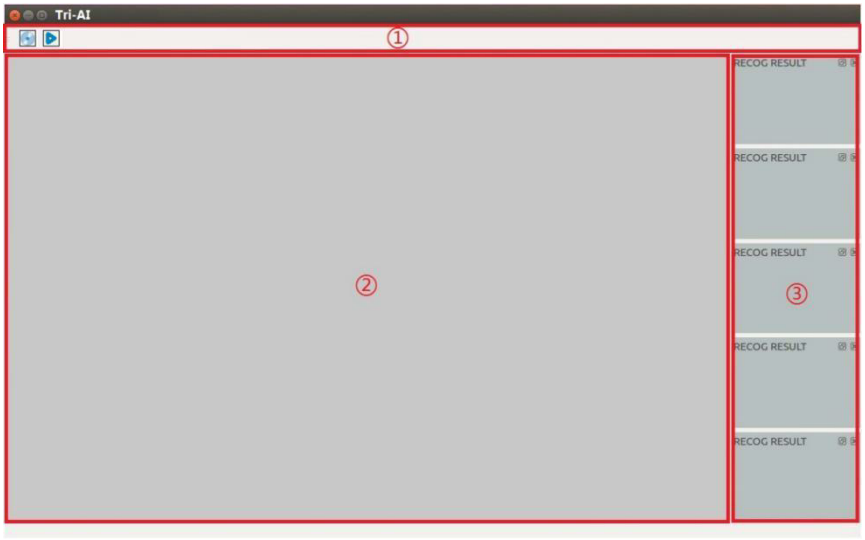

Figure S7. The main page of Tri-AI system, related to Figure 2.

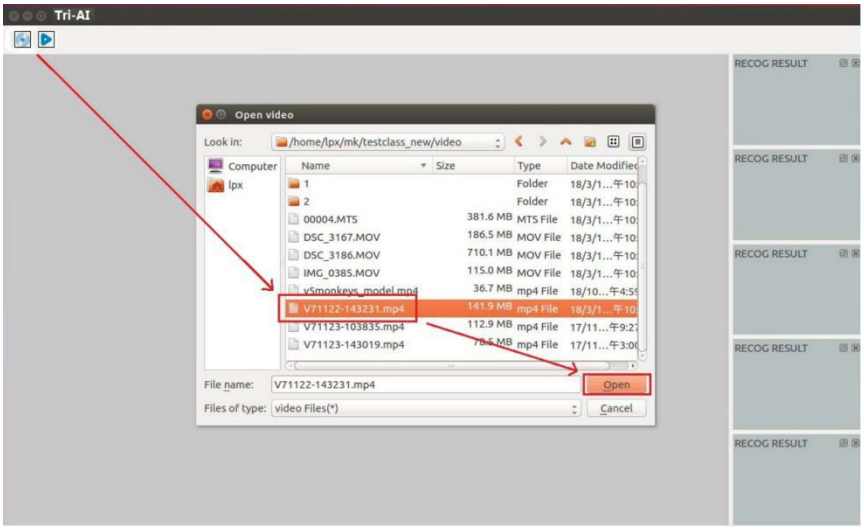

Figure S8. Select an image or a video, related to Figure 2.

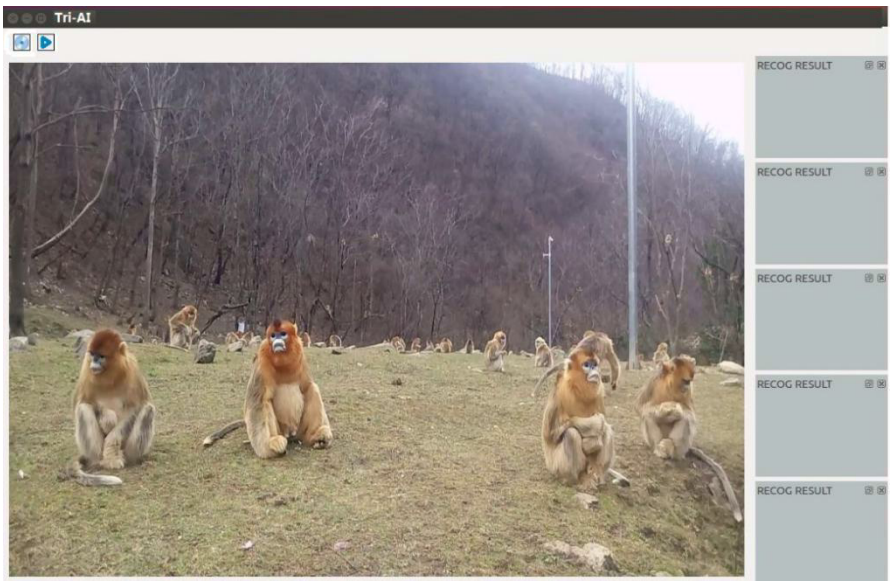

Figure S9. The first frame of a video shown in the area ②, related to Figure 2.

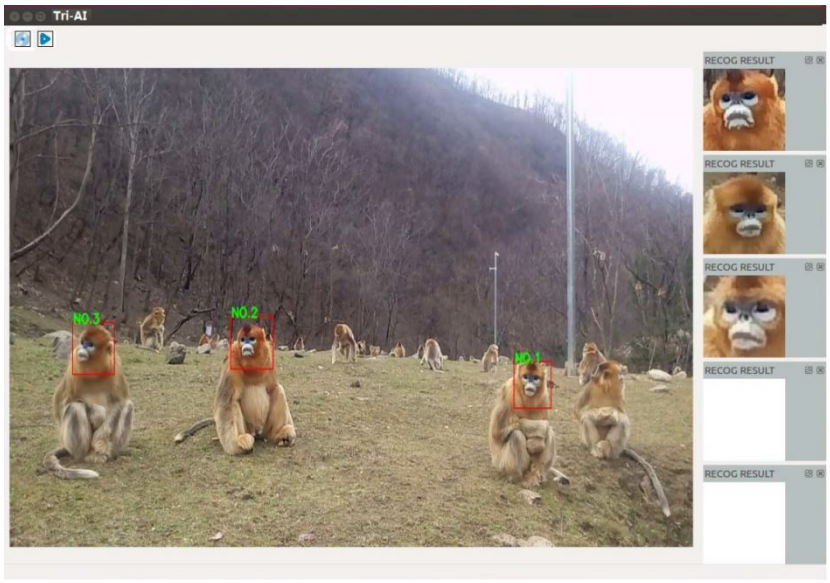

Figure S10. The detection and recognition results, related to Figure 2.

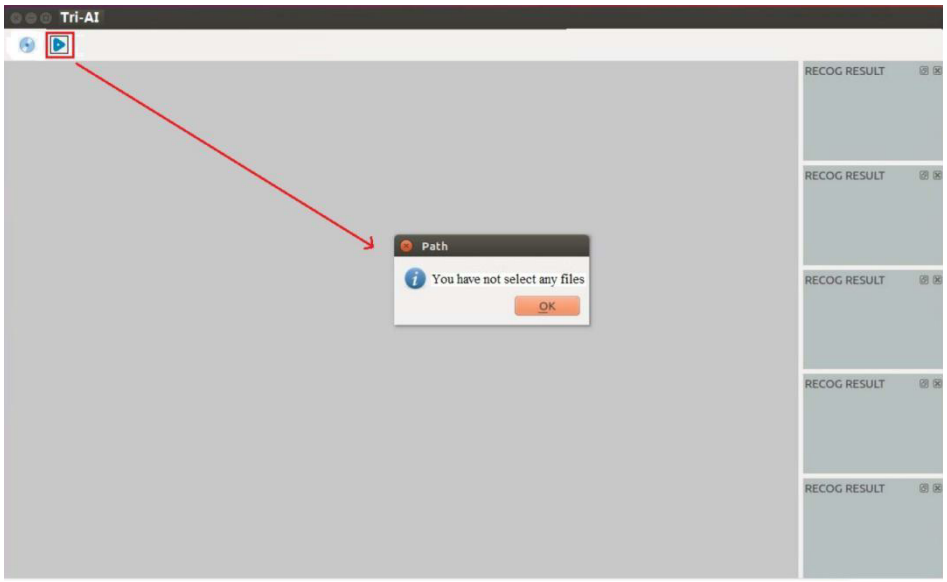

Figure S11. Exception handling, related to Figure 2.

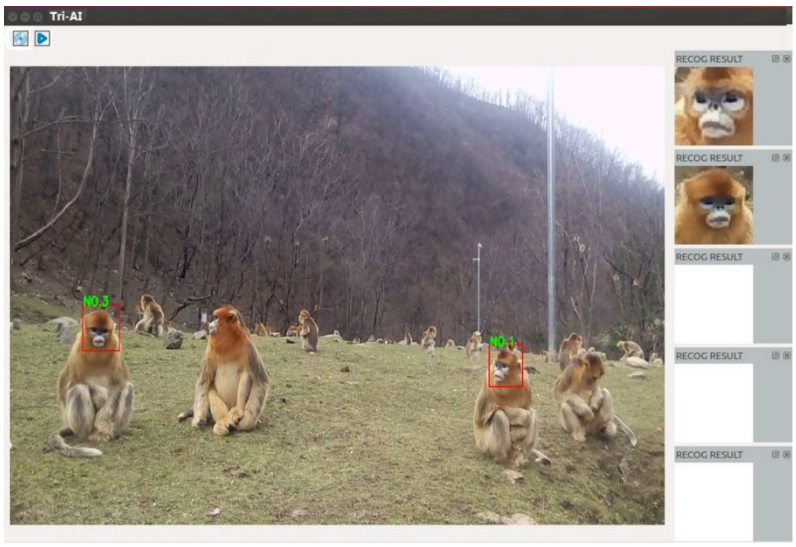

Figure S12. The results of system testing, related to Figure 2.

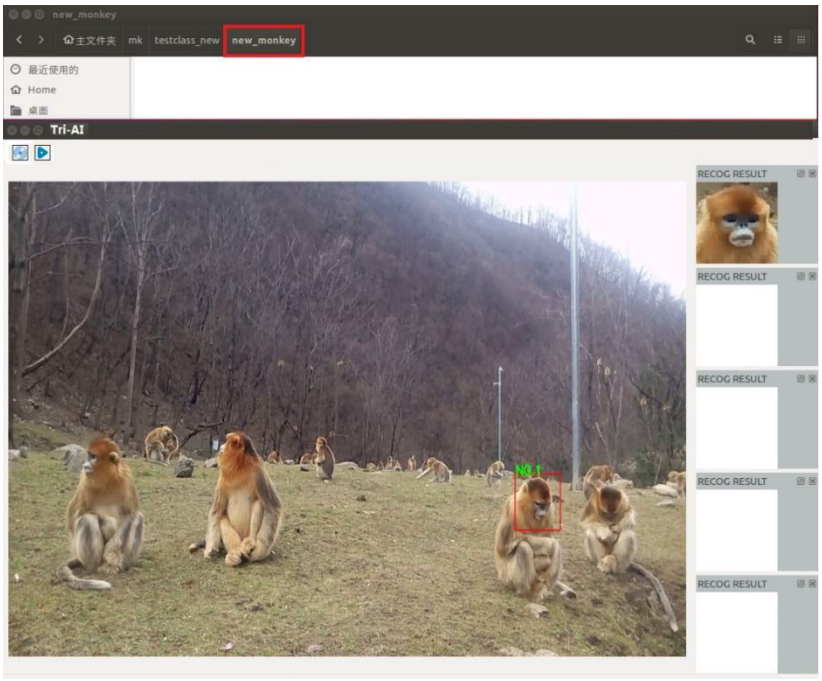

Figure S13. The recognition of a new individual. If the animals have their training facial images in the dataset, their identities can be recognized, related to Figure 2.

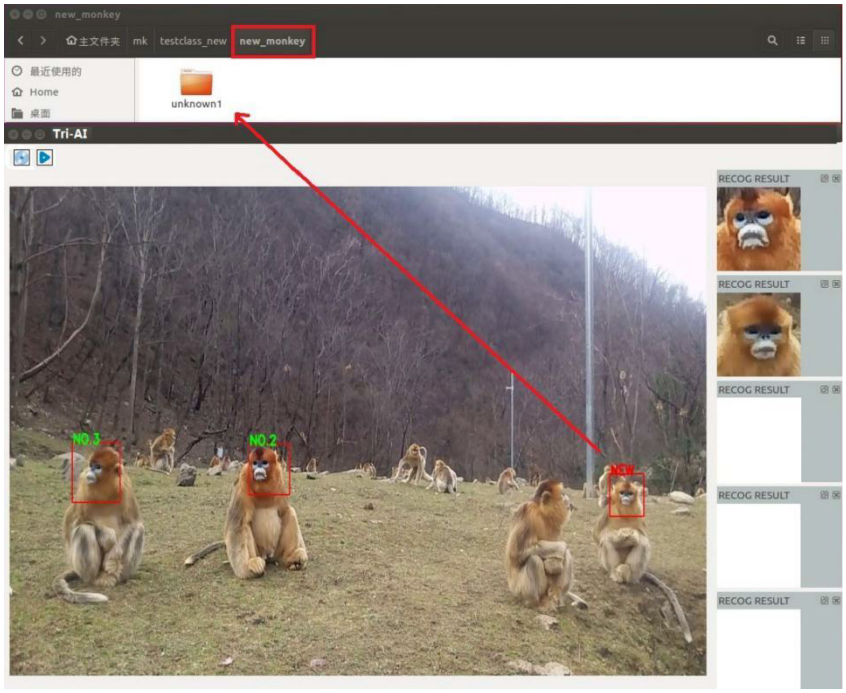

Figure S14. The recognition of a new individual. The new individual is identified, and its facial images are added to the dataset automatically, related to Figure 2.

Figure S15. Usable images of golden snub-nosed monkeys, related to Figure 1.

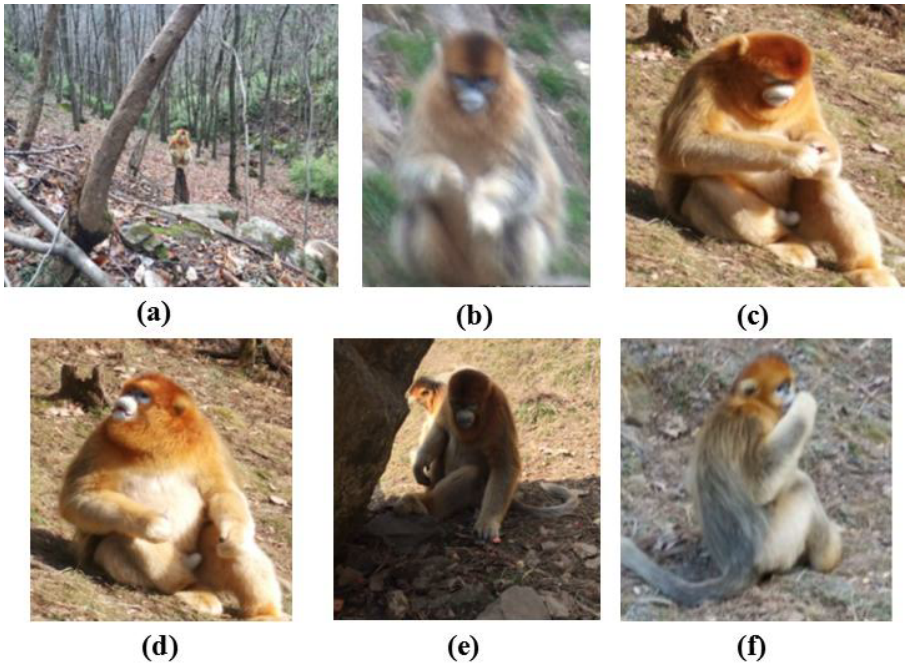

Figure S16. The unusable images under different conditions, related to Figure 1. (a) Too small face. (b) Motion blur. (c) Big angle of the face . (d) Big angle of the face. (e) Under shadow. (f) Covered by its hand.

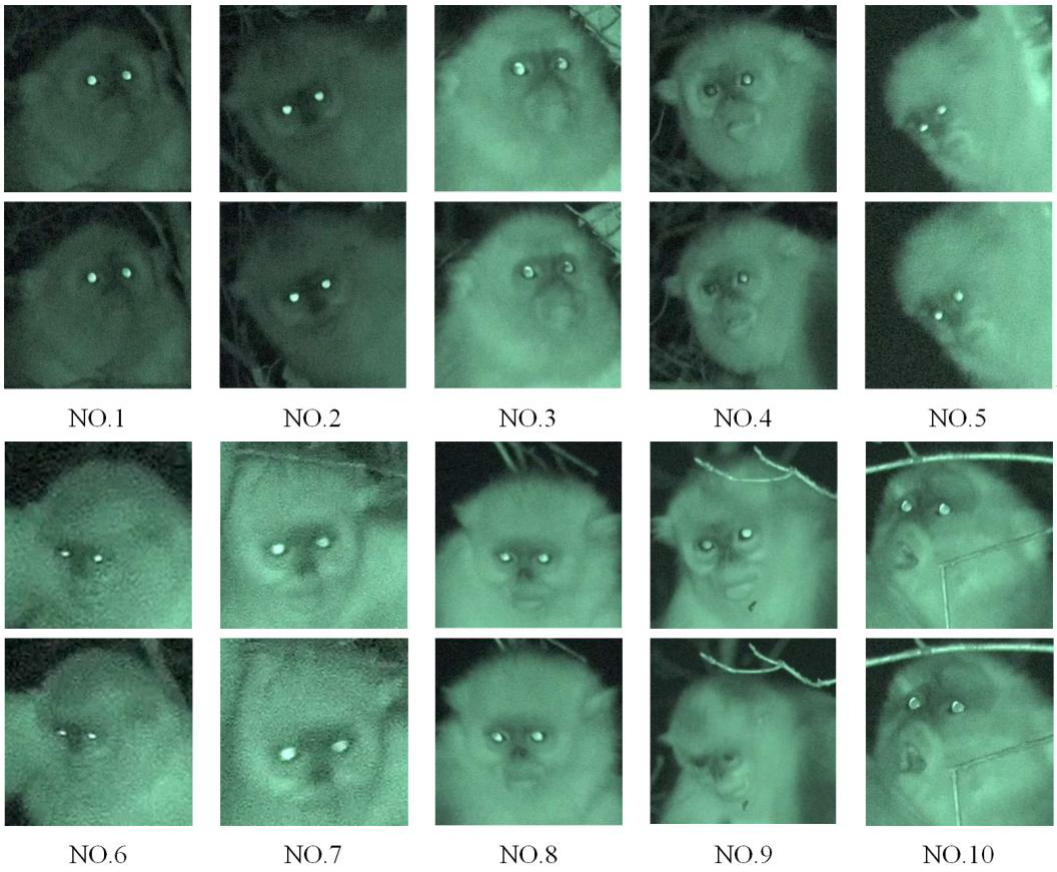

Figure S17. The night facial images of golden snub-nosed monkeys, related to Figures 2 and 3.

S1. Methods

We developed the system Tri-AI using deep learning techniques to automatically detect and identify individual animals by their faces captured from images or videos. Faster Region-based Convolutional Neural Network (RCNN) (Ren et al., 2015) was used to detect the animals' faces and these detected facial images were identified by Tri-attention network, which are described in the following two subsections.

Before Tri-AI could be used, we had to obtain training and testing images and conduct data preprocessing (Related to SI S4). The prepared dataset was used to train Faster RCNN and Tri-attention network to develop the models of face detection and identification. We tested the performances of Tri-AI with a large number of images (Table S1). We have explored the implications of animals' facial image recognition, and the face detection and identification of golden snub-nosed monkeys (*Rhinopithecus roxellana*) with Tri-AI was divided into two major steps.

(1) Face detection by Faster RCNN

To identify individual, we first detected the animals' faces from the images or videos with Faster

RCNN (Ren et al., 2015), which is an effective tool for object detection and has the advantages of simple labeling and high efficiency and accuracy. For object detection, Faster RCNN involves two modules. One module is a deep fully convolution network that provides the proposed regions of the faces, and the other module is a detector to identify the final face regions from the proposed regions. Faster RCNN is a single unified network that is particularly useful for detecting the faces of golden snub-nosed monkeys (Figure S1). After sending the images to a CNN model for feature extraction, the Region Proposal Network (RPN) generates the proposal regions based on the extracted features (Related to SI S6), and approximately 300 proposal regions were generated from each image. These regions were then mapped to the feature map of the last convolution layer of CNN. Finally, the Region of Interest (ROI) pooling layers enabled each ROI to generate a fixed-size feature map and the combination of classification and regression was performed using Softmax Loss and Smooth L1 Loss to locate the animals' faces.

Before detecting the animals' faces, the original parameters in Faster RCNN were trained with sufficient images of the animals. Our labeling task was accomplished by marking their faces on the images with the software of labeling. The trained Faster RCNN could be used to detect the animals' faces from images or videos. For example, we used 1,200 images of golden snub-nosed monkey for labeling to detect the monkeys' faces by Faster RCNN.

To verify the effectiveness of Faster RCNN, we made several test experiments. The trained Faster RCNN was used to detect the faces of golden snub-nosed monkeys, Tibetan macaques and tigers from their images, which were captured by cameras. For the test images, we randomly selected 500 images for each species, and counted the number of the detected faces from the detection results to obtain the detection accuracy of 91.1% for golden snub-nosed monkeys, 97.7% for Tibetan macaques, and 97.7% for tigers. Furthermore, ten videos of 22 golden snub-nosed monkeys were used as the test data to evaluate the success ratio of detection and identification. Then we checked the detection and recognition results frame by frame and obtain 98.70% of face detection accuracy. In addition, we used the full precision-recall curves to evaluate the Faster RCNN detector, and this resulted in some false detection cases.

We tracked animals' faces in the small areas around the locations of their detection in former frames. For a given video, Faster RCNN was used to detect the face regions on the first frame of the video, and then for the following frames. The face regions were detected by Faster RCNN only on

smaller image areas based on the former detected face locations.

(2) Individual identification of primates by Tri-attention network

For fine-grained recognition in the field of image recognition, the deep network models with attention mechanism would focus on a certain part of the image, and the extracted features may be the key features, which have the major differences from those of other classes. Thus, the performance of the classification algorithms can be improved. In this paper, a novel attention network model with three channels (Tri-attention network) was designed for the task of fine-grained individual recognition using primate facial images (Figure S2).

The Tri-attention network has three channels, and each channel has its own attention network models, including object level attention model to extract the global facial features, partial level attention model to focus on the facial region of interest, and significant partial level attention model to select a key smaller facial region for extracting the specific facial features. These three channels with different network structures extract three types of facial features, which are then combined for animal individual identification. Therefore, Tri-attention network focuses on the facial features in different levels simultaneously.

These three models constitute three channels of our Tri-attention network, and each channel pays attention to different areas of the primate facial images at different levels and extracts the corresponding features. The features extracted by each channel are fused by cascade operation and used for classification by softmax.

To prove the validity of the roles of each channel, we conducted ablation experiments on the Tri-attention network. While, the significant partial level attention model is used to extract the fine-grained features on only one small facial region, and these features need to be combined with the global facial feature for individual identification. Therefore, this single stream network is suitable for animal face recognition. We made individual identification experiments using object level attention model and partial level attention model and Tri-attention network. The results show different roles of different channels in Tri-attention network (Table S2). The Tri-attention network has improved performance improvements compared to each single channel.

(3) The Detailed Information of Tri-attention Network

In the Tri-attention network model, the object level attention model mainly deals with global features of the facial image. The partial level attention model is concerned with a relatively fixed local

area in the facial image, and the significant partial level attention model focuses on a specific smaller area selected from the facial images of each individual, with the locations of the areas being different for different individuals.

(A) Object level attention model

The features extracted by convolutional neural network (CNN) contain rich spatial information after multiple convolution and pooling layers. These feature maps need to be transformed into feature vectors by full connection layer. However, some effective information may be lost in this process, and the generalization ability of the model would be reduced. To solve this problem, the class activation mapping (CAM) strategy was used in object level attention model to reduce the impact of the fully connected layer on the loss of feature information. The global average pooling (GAP) layer was used in CAM instead of the fully connected layer in CNN, and the GAP layer calculated the average values of all the pixels in each feature map and converted all these average values into feature vectors for classification. The advantage of global average pooling layer is that there are no parameter settings, which means that the effective feature information can be preserved better, and the risk of overfitting be reduced.

Specifically, CAM was used to replace the fully connected layer with GAP. After GAP, the average value of each feature map in the last convolution layer was obtained. The final features were obtained by a weighted sum. The final extracted features were used as input in softmax for classification. We classified the numbers of regions which had an important impact for individual identification based on CAM. Due to the complex structure of Grad-CAM, we used the basic CAM here by considering the efficiency and complexity of the model.

The network structure of object level attention model has 12 convolution layers, 4 maximum pooling layers, and 1 GAP layer (Figure S3). The combination of the ZeroPadding2D layers and the convolutional layers can keep the sizes of the feature maps unchanged after convolution operations, while the sizes of the feature maps become a half of the original ones only after the maximum pooling operations. Correspondingly, the number of filters becomes twice as the former number after each pooling operation. Since the dimension of the output feature vector is related to the number of feature maps extracted by the previous convolution layer, we have added a convolution layer in front of the global average pooling layer, which ensures that the dimension of the resulting feature vector is consistent with the number of categories.

(B) Partial level attention model

Unlike human facial images, the primates general have more hair on their faces, and the shape and texture information of their hair are easily affected by many factors. However, the facial skin areas are relative invariant with respect to morphology and texture, which makes these skin areas more distinguishable. Therefore, partial level attention model mainly focuses on the facial skin area and minimizes the effects of hair and background on the recognition results. To achieve this, a residual network based on the attention mechanism was used in partial level attention model.

Different from object level attention model, the residual network with attention mechanism is a hierarchical structure. GAP is used to pay attention to extracting the feature maps of the entire facial images in object level attention model, while the residual network with attention mechanism used in partial level attention model is a stackable network structure, which can hierarchically pay attention to specific areas of the facial images. Our designed network model focuses on the skin area in primate facial images.

The structure of residual network with attention mechanism (ResNet-AM) is shown in Figure S4, and it mainly contains two branches, including the trunk branch and the mask branch. The trunk branch is a convolutional neural network consisting of three convolution layers and three ZeroPadding2D layers, and the output feature maps are $T_i(x)$. The mask branch processes the input feature maps to obtain the attention feature maps $M_i(x)$ with the same dimension as $T_i(x)$, and the weights of $M_i(x)$ are normalized. The final characteristics of local areas are expressed as:

$$H_i(x) = T_i(x) \times M_i(x)$$

The partial level attention model is used to focus on the feature extraction of the skin areas. Figure S5 shows the feature map in this model, and one can see this model mainly attends to "skin areas" for feature extraction. This partial level attention model is achieved by maximum pooling layer, which has two functions of reducing the dimensions of the extracted features and removing the redundant features of the images. After maximum pooling layer for down sampling and other layers, the obtained feature maps contain the important features from the original images for individual identification.

(C). Significant partial level attention model

Significant partial level attention model mainly focuses on the most significant area (the key facial

area) of the image, which consists of two steps: area selection and feature extraction. To generate the candidate areas for the facial images, Graph-Based segmentation (GBS) algorithm (Ohayon et al, 2013) are used to divide a facial image into a number of small candidate areas, and the color similarity S between the i^{th} and j^{th} small areas can be calculated with the following equation:

$$S = \sqrt{(r_i - r_j)^2 + (g_i - g_j)^2 + (b_i - b_j)^2}$$

The two areas with the most color similarity were then merged, and the two original sub-areas were removed. This process was performed until there were no original sub-areas left to be merged, so that we could obtain a set of candidate facial areas. For the set of candidate facial areas, we used a trained classification model to choose the key area for each individual, and we chose the merged facial areas with highest classification accuracy as the final key facial areas for the corresponding individual.

The significant partial level attention model can extract the highly separable features of key facial regions, which is reflected in the operation of key facial region selection from a large number of separated facial regions through a trained model (AlexNet) (Krizhevsky et al, 2012). Finally, the selected key facial region for each monkey's facial image was used for feature extraction. After the operation of convolution pooling, only the golden snub-nosed monkey 's eyes and nose were retained (Figure S6).

(D) Feature extraction from the key facial areas.

For the selected key facial areas, VGG16 network is used for feature extraction. VGG16 has 16 convolutional layers, the size of each convolution kernels is 3×3, the convolution step is 1, and the number of convolution kernels increases from the initial 64 to 128, 256, and 512. The size of the convolution kernels in pooling layer is 2×2. VGG16 has three fully connected layers. Since our Tri-attention network consists of three channels, the parameters of the full connection layer of VGG16 would inevitably affect the overall performance of the whole network. Therefore, the parameters of the three fully connection layers are set as 1024, 512, and 128. We used VGG16 rather than residual blocks in the significant partial level attention model to extract the key facial regions by considering the model efficiency.

S2. The operation instruction of Tri-AI

(1) Computer Requirements for Tri-AI

Tri-AI runs on a small workstation with Intel Xeon(R) CUP E5-2650 V4 (2.20GHZ×24), Graphics: GeForce GTX 1080 8g, RAM: 64GB, and Storage: 1TB. The codes of Tri-AI were written in C++ and

Python, and were compiled using g++. Its graphical interface was designed based on Qt, which was developed by the Qt Company in 1991 as an application development framework of cross-platform C++ graphical user interface. The supportive operation system was Ubuntu 14.04 (or an improved version of Ubuntu).

(2) Specific Operations of Tri-AI to Detect and Recognize Faces

Step 1：Start the application. Input the filename of our software in the command line, and press the enter button to start the software. This step intends to accomplish the tasks of animal face detection, identification, and tracking.

Step 2：The main interface will appear when our software finishes loading. The main interface has four areas (Figure S7). The area ① is functional area. Here, one can select the test images or videos and simply click on the "start" button to detect the monkeys' faces from your selected images or videos. The area ② shows the selected images or videos and the corresponding detection and recognition results of monkeys' faces. The area ③ presents the image of the face of the animal in the database that is most similar to the image that the software is being asked to identify.

Step 3：In the area ①, one can click on the first button to open an image file or a video file (Figure S8). When one clicks on "open" button, the selected image or video will appear in the area ② (Figure S9).

Step 4：In the area ①, one can click on "Start" button (the second button), and the system will start to detect the monkeys' faces from the images or the videos and try to recognize who they are. If these monkeys show their frontal faces or profile faces not too far from directly ahead, the individuals will be identified and their corresponding information will be provided (e.g., IDs or names) and an image of each face will be shown in the area ③. These face images are selected from the database, and are the most similar to the monkeys' faces the software is being asked to recognize (Figure S10).

(3) Exception handling

If one did not selected the test images or videos and simply click "Start", a dialog box "You have not select any files" will pop up. In this case, click "OK" (Figure S11), and reselect a test image or video. One then can continue to make the following operations.

(4) System testing

Using golden-snub nosed monkeys as an example, when a monkey's face is directed forward (frontal) or if the animal provides a profile face that is not too far off facing forward, Tri-AI will be

able to recognize this individual (Figure S12). Contrasting the same monkey presenting profile faces with different angles, the system can still get accurate recognition (Figure S10). If animals have their training facial images in the dataset, their identities can be recognized in the images or videos. Otherwise, they are identified as new individuals, and their facial images are added to the dataset automatically (Figures. S13 and S14).

S3. Basic Strategies, Parameters, and Standards for Capturing Animal Facial Images

Compared with acquiring human face images, it is much harder to capture animal facial images because both their living environment and behaviors are much uncontrollable. Therefore, we need to develop specific strategies to acquire different animals' facial images according to their habits and environment. In the process of capturing animals' facial images, it is desirable to use uniform norms and strategies to ensure the animals' facial images obtained within and between studies have high accuracy, regularity, and consistency. Based on our experience, we present guidelines for capturing facial images that are applicable to the theoretical and applied questions researchers are asking and the quality of images that can captured under often difficult field conditions.

(1) The basic strategies for capturing facial images:

(A) If there are few individuals in the population/group being sampled, potentially an observer can see all individuals at once and get independent facial images one by one. This is the ideal situation that can often be done in zoo settings or with very small groups in the wild. However, in most cases, there are many individuals in one area and they cannot be captured in a single image or a set of temporally close images. Therefore, it is necessary for at least two or more people to cooperate, while one taking pictures; others can track the individual's movement to ensure the same individual is not repeatedly photographed.

(B) If there are hundreds of individuals in the population/group or if it is difficult to see all individuals at once as they are somewhat cryptic, it is hard or impossible to separate them artificially, and difficult to distinguish them from memory without potentially months of field work. But if one can temporarily mark individuals, with a harmless color ink for example, then it is possible to mark animals after its photos have been taken. However, all images must be obtained within a short time frame, so that the mark does not fade.

(C) For most species, it will not be possible to mark the animals as they are difficult to see and avoid humans. However, for some group living animals that live in stable small groups, like ring-tailed

lemurs (*Lemur catta*), a number of observers can each select a single or a few animals or a particular area occupied by the group and can quickly take each individuals facial image at the same time. Once Tri-AI is used to assign individuals, it may be possible to study the images to learn the unique individual features.

(D) For species that are cryptic (e.g., many social carnivores) or avoid human observers (e.g., forest elephants (*Loxodonta cyclotis*), lowland gorillas (*Gorilla gorilla*), or giant forest hogs (*Hylochoerus meinertzhageni*)), but repeatedly come to specific locations, such as salt licks or waterholes, it may be possible to set multiple camera traps to simultaneously capture images of all or most of the group.

Above all, it is necessary to ensure that the repeated images, which are taken for the same individual, are unique individuals.

(2) Technical requirements for capturing primates' facial images

It is necessary to develop the basic parameters and standards to ensure that the collected images can meet the necessary requirements and we recommend the following specifications of the equipment, image quality, and quantity.

(A) Equipment: For most mammals, mobile phones take images of sufficient quality when the animals are less than 10 meters from the observer. However, if the animals are farther than 10 meters, high-quality single lens reflex (SPR) cameras which that can take more than 5 images/sec images with over 20 million pixels are recommended.

(B) Image resolution: The primate faces in the images should have 50×50 or more pixels. This is not a difficult standard to meet with most cameras if the animal is relatively close.

(C) Image clarity: The primate faces should be clear, in focus, and not blurred caused by the animals' movement or camera person's hand tremor.

(D) The angles of the face: Images of the face should be taken within the face angles of 30° (or between -15° and +15°) and it is desirable to increase the diversity of the samples from different face angles.

(E) Light: We should avoid taking photos when these animals are in the strong light or under the shadow.

(F) Cover: It is not desirable to have large areas of the animals' faces covered, but if some parts of the face are obscure the images may still be useful.

(G) The number of the images: It is best to have at least one hundred facial images per individual.

(H) Mark the face image data: After the needed number of suitable images are taken, each image must be accurately labeled manually and the mark information should include species, individual identity, age, sex, etc..

(3) Examples of the captured images:

We illustrate the basic parameters, strategies, and standards for capturing facial images using golden snub-nosed monkey.

(A) Usable images: The facial images should be clear, have more than 50×50 pixels, have the face primarily facing forward (i.e., have small side angles), and have appropriate light exposure. The images shown in Figure S15 are suitable for use in our Tri-AI system.

(B) Unusable images. The following images cannot be used by the Tri-Ai System. The monkey's face in Figure S16 (a) is too small, (b) has motion blur, (c) the angles of the monkeys' face is inappropriate as the animal is looking down, (d) the angles of the monkeys' face is inappropriate as the animal is looking to the side, (e) the monkey's face is under shadow, and (f), the monkey's face is covered by its hand.

Using these basic parameters, strategies, parameters, and standards, we were able to capture the facial images of more than forty species of primates. For some of these species, we obtained more photographed more than 1,040 individuals (e.g., golden snub-nosed monkeys in Qinling Nature Reserve and *Macaca thibetanas* in Tangjiahe Nature Reserve). Also, we captured the images of many species of primates from 18 zoos, including those in Beijing, Shanghai, Dalian, Weihai, Qindao, Ningbo, and Shijiazhuang. The images of most golden snub-nosed monkey individuals were captured in a number of different days, while the images of other animals were obtained in single days in the zoos. We got 102,399 facial images of primates.

S4. Primate Facial Image Dataset

Establishing facial image dataset provides an important foundation for the current and future scientific research. To collect sufficient facial images for multiple wildlife individuals, we traveled to 18 zoos in 16 cities and 6 nature reserves in China and took images and videos of 1,040 individuals of 41 primate species, and selected 102,399 facial images.

The collections of these images were fraught with challenges, such animals living in a complex arboreal environment, animals moving very fast, individuals engaging in a wide range of activities,

light conditions changed frequently, and shadow often fell across the animal's face. To acquire these images and videos, we explored many approaches for image acquisition, image selection, facial image extraction, and labeling the images. Based on this experience, we make the following recommendations for future users.

(1) Image acquisition

The acquisition of animal facial images is much more difficult than capturing human facial images, as neither the environment nor the animal's behavior can be controlled. The user community of such approaches needs to develop standardized methods. The basic specifications and related requirements for the acquisition of primate facial images are explained in S3 above. Our early image-capturing work was done with the golden snub-nosed monkey of the Qinling Mountain in China and we collected facial images of 196 individuals. We also travelled to the Tangjiahe Nature Reserve and collected more than 2,000 facial images of 30 wild Tibetan macaques. Further, we captured images from 18 zoos in 16 cities including Chengdu, Taiyuan, Shijiazhuang, Beijing, Qinhuangdao, Dalian, Weihai, Qingdao, Jinan, Shanghai, Ningbo, Hangzhou, Suzhou, Wuxi, Nanjing, and Xi'an. In total, we obtained 102,399 facial images of 1040 individuals, from 41 species.

(2) Image data processing

After obtaining the images, the facial areas of each primate individual must be extracted from the images using manually screenshot methods or face detection methods, such as Faster RCNN. Primates are gregarious animals, which typically result in individual images showing multiple individuals. In this case, the most challenging task for us is individual recognition as many animals look so similar. Therefore, we had to carefully identify which animal the facial image was from. We did this by checking the position of each in all the images or repeatedly comparing the differences between them through visual interpretation. Therefore, most of the facial images were identified manually using screenshots to avoid data confusion. This manual method of face detection can only be applied to those images which have one or very few individuals. We found it best to have all the facial images in our dataset to be square, contain almost all the facial information of the individuals, and have as little background as possible.

(3) Construction of Primate Facial Image Datasets

When image acquisition and processing are completed, researchers need to build a dataset of facial images. For our dataset, we classified all the facial images by species and the facial images of

each individual ware assigned a unique label. In our data set, there were differences in the number of individuals among species (e.g., golden snub-nosed monkey had 43,304 facial images of 227 individuals and each individual had an average of 191 facial images, *Cebus apella* had 3,026 facial images of 43 individuals, and the average number of facial images for each *C. apella* individual was 70). The average number of facial images for each individual in our dataset was 99. Finally, our image dataset had a total of 1,040 individuals and 102,399 facial images. A detailed list of all primate species, the individual numbers in each species and the total number of facial images of each species in our dataset are given in Table S1.

Our images of golden snub-nosed monkeys were captured in the wild. Thus there was greater variance in the images, compared to the other species where images were obtained from zoos, typically on single days. The dataset has been made publicly available at the database:

(AFD(Animal Face Database): http://dx.doi.org/10.17632/z3x59pv4bz.2).

S5. Identification of golden snub-nosed monkeys using taken at night

In total, 581 facial images taken of 24 golden snub-nosed monkeys at night were used for face recognition. Here, 60% were used as training samples, 10% for validation, and 30% as test samples. The identification accuracy of these night images was 92.03% and the night vision images for 10 golden snub-nosed monkeys are shown in Figure S17.

S6. A list of technical terms

CNN: A convolutional neural network (CNN or ConvNet) is a class of deep, feed-forward artificial neural networks that have successfully been applied to analyzing visual imagery. An early development of CNN for facial recognition was developed by Lawrence and colleagues (Krizhevsky et al, 2012), but it has been subsequently improved (Lecun et al, 2015).

RCNN: Regions with CNN features. RCNN is an object detection model based on the CNN network. It uses the selective search method to get 2,000 candidate boxes with different sizes, and then CNN network is used to extract the regional features for the classification of the objects and background.

Fast RCNN is a fast object detection model based on multi-task learning. In the training phase, Fast RCNN is 9 times faster than RCNN. During the testing phase, Fast RCNN is 213 times faster than RCNN (Girshick, 2015).

RPN: Region Proposal Network. RPN can obtain a series of object proposals from arbitrary images. It provides the suspected areas for object detection models.

Faster RCNN: This is a newer version of an object detection model of fast RCNN. The network structure mainly includes RPN and fast RCNN. RPN is used to select the suspected areas where the objects may exist in the image. The fast RCNN is used to identify whether these suspected areas actually are the objects.

ResNet: Residual Network. The residual network uses the convolutional layers to perform residual learning to solve the problem of performance degradation when the network goes deeper. The existing ResNet models have ResNet-20, ResNet-34, ResNet-51, ResNet-101, ResNet-152, and other improved related network models.

Shallow Resent: Shallow ResNet is an improved deep residual network proposed in this paper. Shallow ResNet simplifies the network structure based on the traditional ResNet. The convolution layers are increased to form new types of residual blocks, which can improve feature learning ability. Shallow ResNet is a good solution to the problem that the traditional network models with fewer layers are hard to extract the deep features of the animal facial images, but the features extracted by deep network models lose more information.

Supplemental References

Girshick, R. (2015). Fast r-cnn. IEEE international conference on computer vision (ICCV) (IEEE, New York). 1440-1448.

Ren, S., He, K. M., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NIPS). 91-99.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Conference on Neural Information Processing Systems (NIPS) (Curran Associates Inc). 1097-1105.

Lecun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. Nature, 521, 436-444.

Ohayon, S., Avni, O., Taylor, A. L., Perona, P., and Egnor, S. R. (2013). Automated multi-day tracking of marked mice for the analysis of social behaviour. J Neurosci Methods, 219, 10-19.