Detecting and Tracking Bottoms and Faces of the Crested Black Macaque in the Wild

John P. Chiverton[1]
john.chiverton@port.ac.uk
Jerome Micheletta[2]
jerome.michelatta@port.ac.uk
Bridget M. Waller[2]
bridget.waller@port.ac.uk

[1] School of Engineering, University of Portsmouth, UK.
[2] Centre for Comparative and Evolutionary Psychology, University of Portsmouth, UK.

Abstract

Monkeys are important to many areas of science and ecology. The study of monkeys and their welfare are important components requiring complex observational studies. This work is therefore concerned with the development of computer vision techniques for the purposes of detecting and tracking monkeys with the ultimate aim to help in such studies. Monkeys are complex creatures for the purposes of tracking because of complex deformations. This complexity is further compounded by an *in the wild* setting where forest conditions result in frequent occlusions and changes in lighting. Despite these complexities monkeys present some interesting features that can make detection and tracking possible: their bottoms and faces. A system is thus described consisting of detectors trained to detect faces and bottoms of monkeys which are used within a tracking framework to initialise a system of tracklet construction. Steps are also described to enable disparate but coincident tracklets to be merged thus enabling longer run analysis of individual monkey movements.

Experiments are performed using image data taken from video footage of Crested Black Macaques in natural forest surroundings. Results demonstrate relatively successful detection of monkey bottoms where the correspondence analysis and tracking process helps to reduce false positives.

1 Introduction

The aim of this work is the development of computer vision techniques applicable to the detection and tracking of monkeys in real-world settings (*in the wild*). These techniques aim to support the work of animal behavioural psychologists, evolutionary psychologists and people interested in the welfare of monkeys.

Monitoring of monkeys is often performed to assist in studies of social interaction such as huddling, social play, grooming, aggressive interaction. This monitoring process is usually performed manually with people observing the monkeys, either live or through recorded material [14]. This can be time consuming, labour intensive and expensive. Furthermore it can be difficult to observe all the activities of all the monkeys all of the time. Therefore it could potentially be useful to include automation in the monitoring process [5].

Automated monitoring of monkeys has included an outdoor 3-D visual tracking system for the study of spatial navigation and memory in rhesus monkeys in [9]. Khan *et al*. used a pair of cameras to extract 3-D trajectory information including path length and speed information. This was used to determine spatial navigation abilities of the monkeys. This work was applicable to outside environments however it was based on background subtraction, requiring stationary cameras, continual movement of the monkey and only a single point in 3-D was extracted. Furthermore the automated aspect was just limited to the extraction of single trajectory like information rather than activity information for multiple monkeys.

Later work [7] included a fuller 3-D solution to the observation of monkeys using a 4 camera system in an indoor enclosure. The described system included a monkey detection stage using a linear SVM with supervised training of monkey images. The classifier was used for the classification of regions of the images to detect the locations of the monkeys. Background subtraction was then used to determine accurate silhouette information on the bounds of the locations of the monkeys. This work was limited to an indoor scene. Furthermore the background subtraction assumed the background did not require updating because the enclosure was indoors and within a highly controllable environment.

Another work [1] was based on colour detection, requiring the monkeys to wear uniquely coloured collars, often used for restraining. This work was also limited to an enclosure. The use of colour coded collars also makes the vision problem simpler but it is not applicable to tracking monkeys in real world environments. A further advantage is that it enables tracking in frames consisting of occlusions and other difficulties such as changes in appearance or similar.

Monkeys, like many wild animals have an appearance that typically prevents easy delineation from the natural surrounding environment such as a forest. However, like many other types of wild animals, many types of monkeys also possess some anatomy that is distinctive and relatively easy to spot. For the Crested Black Macaque, it is their bottoms which are pink or even red for both males and females. Some examples images can be seen in Fig. 1. For comparison, their faces have fewer distinct features. We will investigate the use of both monkey faces and monkey bottoms for the purposes of detection and tracking as these can both provide important information on the location, tracking and social interaction of individuals. This work will apply a number of feature extraction and classification techniques for the purposes of monkey bottom and monkey face detection. Histograms of Oriented Gradients (HOGs) as described in [6] in conjunction with linear Support Vector Machine (SVM) will be considered. HOGs were originally described for person detection although they have been used in a simplified static camera, whole monkey detection system for indoor enclosure monitoring by [7]. Local Binary Patterns (LBP) in conjunction with a cascade of classifiers as described in [10] will also be considered here for application to the detection of monkey faces and monkey bottoms which are more famously used for human face detection and recognition.

First we will discuss the tracking framework and then in the following section on experiments we will describe experiments associated with the different detection approaches in conjunction with the described tracking framework.

2 Methodology

Monkeys are initially detected using distinctive features and then these detections are processed as part of a tracking framework.

At discretely sampled time instances $t \in \mathbb{N}$, the distinctive features are defined here to

Figure 1: Top row: selection of the 824 hand segmented rear views of the Crested Black Macaque's bottoms that were used for feature vector training. Bottom row: Selection of the images of Crested Black Macaque's faces used for feature vector training from the manually segmented 186 frontal and 512 side views of monkeys' faces.

occupy the majority of a bounded rectangle $R^{[t]}$; which for convenience are defined as sets of pixel locations $(x,y) \in R^{[t]} \subseteq \Omega$ where $\Omega \subseteq \mathbb{N}^2$ is discretely sampled image space defined on the set of two dimensional natural numbers. For a particular frame in time t there may be multiple detections, i.e.

$$\mathbf{R}^{[t]} = \left\{ R_1^{[t]}, R_2^{[t]}, ..., R_N^{[t]} \right\}. \tag{1}$$

where $R_i^{[t]}$ are the bounded rectangles. A distinctive region $R_i^{[t]}$ can be tracked across frames $R_j^{[t+q]}$, $(q \in \mathbb{N})$ where correspondence is defined to occur between regions $R_i^{[t]}$ and $R_j^{[t+q]}$. Correspondence is estimated here by calculating the similarity of the regions in proximity and appearance. This process is often error prone due to changes in shape and appearance because of *e.g.* changes in lighting, particularly in video frames from forest scenes and other *in the wild* settings. For example, the detection process will often result in false positives (FPs). Many FPs can be filtered out by a correspondence and tracking step between frames. Determination of correspondence between frames can help to remove FP detections.

Correspondence is estimated here based on feature vectors F_i and F_j associated with regions R_i and R_j respectively. A measure of distance between feature vectors is given by the exponential of the normalised Euclidean distance

$$e_{i,j} = \exp\left(-\frac{1}{2}(F_i - F_j)^T \Sigma^{-1} (F_i - F_j) \right), \tag{2}$$

where Σ is a diagonal covariance matrix of the feature vectors and is assumed to be a diagonal matrix for the purposes of the present work resulting in (2) taking the form of a Gaussian kernel. Other ways of measuring the disparity between regions such as looking at the effect of a non-diagonal covariance matrix could also be considered resulting in a kernelized Mahalonobis distance. Similarly other distance measures could be investigated such as the Earth Movers Distance (EMD) or Hausdorff but (2) is used here for this present work.

For correspondence matching it is desirable to find the set of detections in time frame t that are most similar to the set of detections in other time frames such as $t+q$ which

maximise $e_{i,j}$. Considering a potential correspondence between a single pair of regions $R_i^{[t]}, R_j^{[t+q]}$ one could find $c(j) = \text{argmax}_{i:\forall k \neq i, R_k^{[t]} \in \mathbf{R}^{[t]}} e_{i,j}$. However it does not consider the entire set of possible matches which may result in a more optimal set of correspondences. Considering all the regions in the two time frames we can therefore consider

$$L = \max_{Y} \left[\sum_{\forall j: R_j^{[t+q]} \in \mathbf{R}^{[t+q]}} e_{y_j,j} \right] \tag{3}$$

where the sequence

$$Y = \left(y_j \in \left\{ i \middle| R_i^{[t]} \in \mathbf{R}^{[t]} \right\} \cup \{\varepsilon\} : \left(1 \leq j \leq \left| \mathbf{R}^{[t+q]} \right| \right); (y_j \neq y_k, \forall\, k \neq j, \text{if } y_j \neq \varepsilon) \right) \tag{4}$$

potentially contains empty strings together with unique elements as indicators to regions from time frame t. The empty string is for the case of when there is no correspondence with the preceding frame where $e_{\varepsilon,j}$ takes a fixed cost that penalises no correspondence. The maximisation in (3) by manipulation of the mappings in (4) is a complex problem to solve if considering a greedy search. However [11] suggested using the Singular Value Decomposition (SVD) of the affinity matrix E with elements $e_{i,j}$ by computing $E = USV$ where U and V are matrices of the orthonormal eigenvectors and S contains the singular values. The matrices U and V together with setting the singular values of S to 1 (forming B) can then be used to determine a new matrix $G = UBV$. This new matrix will now contain row to column maximums for correspondences that maximise (3). Similar approaches have been used since *e.g.* [11], then developed by [12] and is still being developed for more sophisticated arrangements, see *e.g.* [2].

A series of matched correspondences over a period of time, $[t, t+k]$ can form a tracklet where a detected item has been successfully tracked over a finite period of time. These tracklets may not cover the entire existence of the object in the entirety of the image data *e.g.* due to occlusions. Techniques for the matching and merging of tracklets are therefore included here.

Kalman filter state estimation individualised for each detection with a found correspondence is used to forward-propagate and back-propagate by $\pm v$ time instances. The advantage of this approach is the explicit consideration of the correspondence matching process. Other tracking based approaches could also have been used such as a mixture based approach, *e.g.* [13] that allow a more elegant solution to the problem of tracking multiple targets. However other tracking techniques often do not consider the problem of explicit tracklet formation and in particular tracklet propagation across time both forwards and backwards. This allows disparate tracklets to merge across severe occlusions and other potential problems that may occur at the detector level such as due to significant changes in lighting.

Forward-propagation propagates a tracklet beyond a detection at a particular time instance $t_1 + k_1 + v$: $[t_1, t_1 + k_1 + v]$ and back-propagation propagates the tracklet backwards $[t_1 - v, t_1 + k_1]$. This means that tracklet

$$K_1 = \left\{ R^{[t]} \middle| \left(R^{[t]} \in \mathbf{R}^{[t]} \right), (t \in [t_1, t_1 + k_1]) \right\} \tag{5}$$

after propagation can be estimated, and defined here, to potentially exist over the range $[t_1 - v, t_1 + k_1 + v]$:

$$K_1^{\pm} = \left\{ R^{[t]} \middle| (t \in [t_1 - v, t_1 + k_1 + v]) \right\}, \tag{6}$$

Figure 2: Selection of bounded images used for negative feature vector training. Here 427 images from forest scenes were hand-labelled for the presence of monkeys and then random sampling in terms of location and scale was performed to extract out 6829 images.

where $R^{[t]} \in K_1$ for $t \in [t_1, t_1+k_1]$; $R^{[t]} \in K_1^+$ for $t \in [t_1+k_1+1, t_1+k_1+v]$ where K_1^+ is the set of forward propagated regions for tracklet K_1; and $R^{[t]} \in K_1^-$ for $t \in [t_1-v, t_1-1]$ where K_1^- is the set of backward propagated regions for tracklet K_1.

A Kalman filtered update step is used in the absence of measurement information across the propagated time instances for the region being tracked (a rectangle for our particular implementation). The existence of the tracklet at these propagated time instances is tentative. A further tracklet at a future time instance K_2 that has also been propagated $K_2^\pm$ with $[t_2-v, t_2+k_2+v]$ is a potential candidate for merging if the two tracks intersect during the propagated range. Formally, a new super-tracklet $K' = K_1 \cup K_2$ can be formed if the tracks intersect $K_1^\pm \cap K_2^\pm \neq \emptyset$ where the propagated time instance ranges are $[t_1-v, t_1+k_1+v] \cap [t_2-v, t_2+k_2+v] \neq \emptyset$ and $[t_1, t_1+k_1] \cap [t_2, t_2+k_2] = \emptyset$. Super-tracklet K' has range $[t_1, t_2+k_2]$, assuming tracklet K_2 occurred after tracklet K_1, i.e. $t_2 > t_1+k_1$ or $[t_2, t_1+k_1]$ for the opposite case. This tracklet merging and propagation process helps to overcome problems such as occlusions and certain limitations associated with the Kalman filter because it enables quite different motion to be associated with different sub-ranges of time but for the same tracklet (after merging).

The system described has been implemented in C++ with the use of OpenCV [3]. The entire framework is also made freely available on a popular repository hosting service for open source code [4]. Experiments using the system are now described.

3 Experiments and Results

High Definition video footage of Black Crested Macaques in forest locations is used here for testing the proposed system. Individual frames from a number of 1 to 2 minute sequences were extracted and hand labelling was performed on: 824 rear views of monkey bottoms; 438 side views of monkey bottoms; 186 front views of monkey faces; and 512 side views of monkey faces. These were used as positive training samples which were cropped to a size of 128×96 pixels. Example positive images used for training the classifiers can be seen in Fig. 1. Negative samples were generated by random sampling and scaling from the background of 427 hand labelled image sequences where the presence of monkey(s) were also manually identified. This generated 6829 samples without any monkeys being present. Example negative background images can be seen in Fig. 2. All results are obtained from non-training data, *i.e.* videos which were not used as part of the training process. Exemplar frames from tracking the rear view of the monkey's bottoms can be seen in Fig. 3. Here

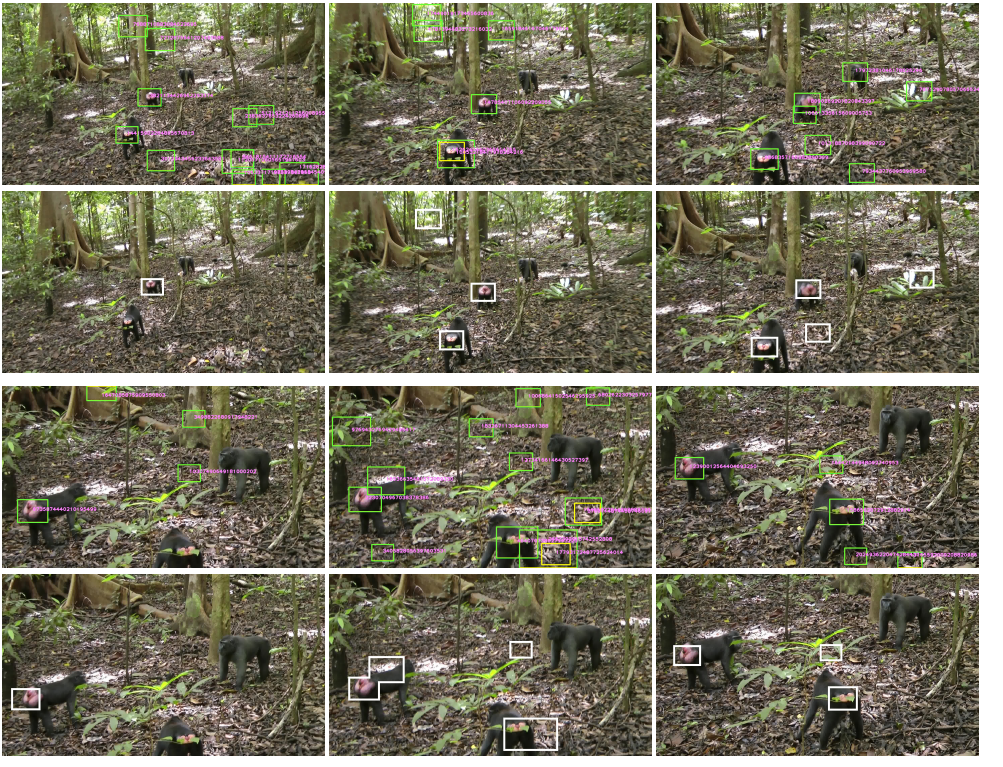

Figure 3: Detection results obtained for tracking the rear view of bottom. 1st and 3rd rows show the detection results before tracking. 2nd and 4th rows show corresponding detection results after tracking.

the initial detections in the absence of the correspondence and tracking part are compared with the results obtained after correspondence and tracking. Most of the false positives have been removed whilst most of the true positives have been retained. This is true for nearly all frames in this sequence as can be observed by the results shown in Fig. 4. For the detection and tracking of the monkey faces, exemplar results can be seen in Fig. 5. Results demonstrating the number of false positives being removed can be seen in Fig. 4.

The mean sensitivity for the HOG based detection of the monkey bottoms was found to be 92% which decreased to 77% after correspondence analysis and tracking. This is in contrast to the precision, which increased from 29% to 59% after correspondence analysis and tracking was introduced. For LBP based detection of the monkey faces, the mean sensitivity was found to be 96% which reduced to 65% after introducing correspondence analysis and tracking. The sensitivity increased from 6% to 23%.

4 Discussion and Conclusions

Receiver Operator Characteristic (ROC) based curves could not be generated here for the tracking results because the tracking process does not result in False Negatives (FNs) preventing the calculation of the False Positive Rate (or $1-$ Specificity). However, the topic of on-going work is the effect of training size and other key parameters on the detectors' performances. An important point to note here is that the training of the detectors was performed

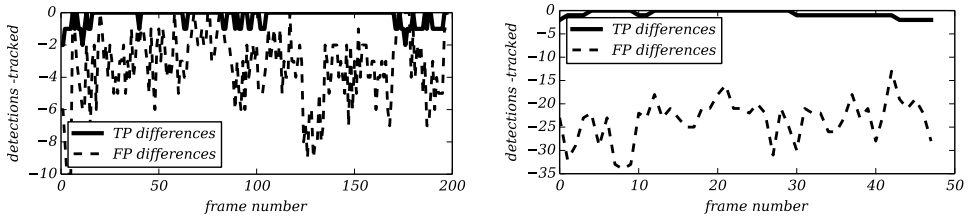

Figure 4: Results from comparing the true positives and false positives for the (left) HOG based tracking of the monkey bottoms and the (right) the LBP based tracking of the monkey faces. These values were obtained by subtracting the number of TPs or FPs raw detections with the number of TPs or FPs after correspondence analysis and tracking. Here it can be seen that many false positives have been removed whilst retaining the original number of true positives for the majority of cases.

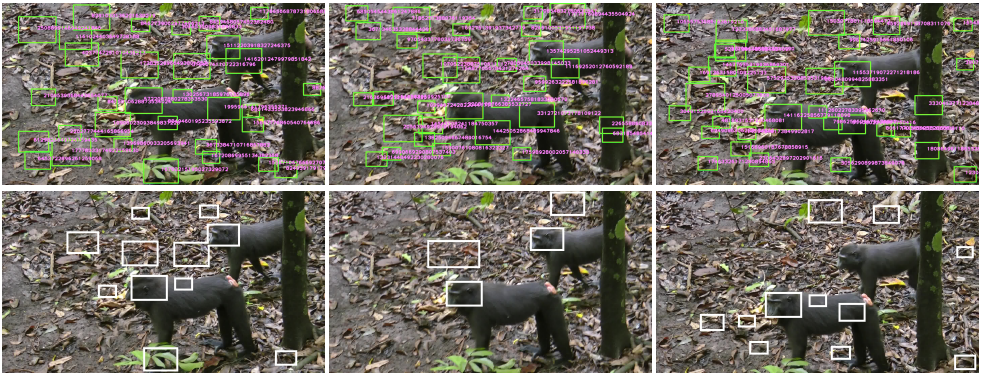

Figure 5: Detection results obtained for tracking the faces. Top row shows the detection results before tracking. Bottom row shows the detection results after tracking.

on video data that was not used for the testing of the tracking.

This work has looked at the detection and tracking of monkeys using distinctive features such as their bottoms and faces. Promising results have been demonstrated for the detection and tracking of the monkey bottoms however more work is required to improve the monkey face detection process, to make it more precise and more sensitive. Only 186 frontal views and 512 side views of monkey faces were used in the training of the LBP based cascade. Many more images have been used in the training of other types of detectors such as those for pedestrian detection, see *e.g.* [6]. In terms of the overall system, the robust detection of both monkey faces and monkey bottoms could be used in behaviour analysis of monkeys whilst being potentially useful for scientific studies and welfare monitoring.

Acknowledgements We thank the Indonesian State Ministry of Research and Technology (RISTEK), the Directorate General of Forest Protection and Nature Conservation (PHKA) and the Department for the Conservation of Natural Resources (BKSDA), North Sulawesi, for permission to conduct research in the Tangkoko Nature Reserve, from which the audio-visual material used in this study were derived. We thank all members of the Macaca Nigra Project for their support.

We also thank the makers of OpenCV [3] and SVM light [8].

References

[1] S. Ballesta, G. Reymond, M. Pozzobon, and J.-R. Duhamel. A real-time 3D video tracking system for monitoring primate groups. *J. Neuroscience Methods*, 234:147–152, 2014.

[2] M. Bansal and K. Daniilidis. Joint spectral correspondence for disparate image matching. In *CVPR*, pages 2802–2809. IEEE Computer Society, 2013.

[3] G. Bradski. The OpenCV Library. *Doctor Dobb's Journal of Software Tools*, 25(11): 120–126, 2000.

[4] J. Chiverton. ggmoda14: A framework for object tracking. https://github.com/chivertj/ggmoda14, 2015.

[5] S. Cooke, S. Hinch, M. Wikelski, R. Andrews, L. Kuchel, T. Wolcott, and P. Butler. Biotelemetry: a mechanistic approach to ecology. *TRENDS in Ecology and Evolution, 19(6)*, pages 334–343, 2004.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893. IEEE, 2005.

[7] N. Ghadar, X. Zhang, K. Li, D. Erdogmus, G. Thibault, A. Bayestehtashk, I. Shafran, K. Coleman, and K. Grant. Visual hull reconstruction for automated primate behavior observation. In *IEEE Int. Workshop on Machine Learning for Sig Proc.*, 2013.

[8] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-scale SVM learning practical. MIT-press, 1999.

[9] Z. Khan, R. Herman, K. Wallen, and T. Balch. An outdoor 3-d visual tracking system for the study of spatial navigation and memory in rhesus monkeys. *Behavior Research Methods*, 37(3):453–463, 2005.

[10] S. Liao, X. Zhu, L. Zhang, and S.Z. Li. Learning multi-scale block local binary patterns for face recognition. In *International Conference on Advances in Biometrics (ICB'07)*, pages 828–837. Springer-Verlag, 2007.

[11] G. Scott and H. Longuet-Higgens. An algorithm for associating the features of two images. *Proceedings: Biological Sciences*, 244(1309):21–26, 1991.

[12] L.S. Shapiro and J.M. Brady. Feature-based correspondence - an eigenvector approach. *Image Vision Comput.*, 10:283–288, 1992.

[13] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multi-modality through mixture tracking. In *Ninth ICCV*, volume 2, pages 1110–1116. IEEE, 2003.

[14] C. Young, B. Majolo, O. Schulke, and J. Ostner. Male social bonds and rank predict supporter selection in cooperative aggression in wild Barbary Macaques. *Animal Behaviour, 95*, pages 23–32, 2014.