A new approach to detecting and analyzing monkey activities using the improved YOLOv5

Biao Shen, Zhiyi Zhang, Qishun Wang, Zhengzheng Tu, Ning Liu

Biao Shen, Zhiyi Zhang, Qishun Wang, Zhengzheng Tu, Ning Liu, "A new approach to detecting and analyzing monkey activities using the improved YOLOv5," Proc. SPIE 12636, Third International Conference on Machine Learning and Computer Application (ICMLCA 2022), 126361V (25 May 2023); doi: 10.1117/12.2675184

Event: Third International Conference on Machine Learning and Computer Application (ICMLCA 2022), 2022, Shenyang, China

A New Approach to Detecting and Analyzing Monkey Activities Using the Improved YOLOv5

Biao Shen[1, 2, a], Zhiyi Zhang[3], QishunWang[1], Zhengzheng Tu[1,b(*)], Ning Liu[2,3,c(*)]
[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China
[3] State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
[a]1633696485@qq.com
[b]zhengzhengahu@163.com
[c]liuning@ibp.ac.cn

Abstract

With the development of science and technology, video equipment has become popular in laboratories. Researchers using animals (e.g., macaques) as models often collect a large number of animal behavior videos for the research aims. Researchers often need to watch numerous videos to record and analyze animal movements and behaviors for subsequent analysis. Manual analysis of videos is time-consuming and inefficient. To resolve this problem, we developed an approach to detect macaques in video and then capture information about their location. Using this information, we can quickly analyze animal behaviors, such as movement preferences and behavior patterns. The method was developed based on the YOLOv5 algorithm and had excellent accuracy and robustness. In addition, this method can automatically process videos and generate statistical results and analysis charts. Furthermore, all settings are manually configurable and easy to use to meet the different requirements of researchers. We believe that this new approach will benefit many researchers.

Keywords: Deep Learning Technique (YOLOv5), Detection, Monkeys, Auto Motion Analysis, Intelligent Video Understanding

1 Introduction

With the widespread use of video equipment in the laboratory, animal research no longer just records animal behavior with pen and paper but also uses video to make more extended and comprehensive records of animal behavior. Manual analysis is relatively routine. Researchers can always watch videos one by one with the naked eye. Such work is time-consuming and inefficient. Moreover, this manual method of scanning videos and recording data is susceptible to the experience and state of the viewer. For example, inexperienced or tired viewers may miss some video information or record incorrect data, affecting the data's reliability. Therefore, there is an urgent need for current research on a new approach that can overcome the shortcomings of the manual video analysis methods described above to meet the needs of laboratories, drug evaluation organizations, etc. Several studies have already developed new techniques to address these issues. For example, the video-based analysis of macaque movement [1,2,3,4] (to quantify overall movement intensity), movement trajectories [2,5,6,7,8,9,10,11] (to measure the trajectory of the body during movement), and behavioral classification [12,13] (to classify different types of activities) have been developed. These approaches can provide valuable information for various experimental purposes. For example, trajectories reflect not only the overall activity level of the animal but also important spatial information about its movements [2]. Due to its development, deep learning has also been widely used in video tracking. An increasing number of deep learning-based methods are used in video analysis [5,6,7,9,10].

Deep learning-based trajectory tracking methods have benefited from the development of deep neural networks. In recent years, the introduction of deep learning has led to the emergence of many high-performance methods in animal tracking [7,9]. Most methods use pose estimation to track key points [5,7,9] or object detection models to track the whole animal [6,10]. High-dimensional motion information can be obtained by recording the motion trajectory of key points or the entire

Third International Conference on Machine Learning and Computer Application (ICMLCA 2022),
edited by Shuhong Ba, Fan Zhou, Proc. of SPIE Vol. 12636, 126361V
© 2023 SPIE · 0277-786X · doi: 10.1117/12.2675184

animal [14].

Although object detection models, such as YOLOv5 [15], and Faster R-CNN [20], have achieved great accuracy in object detection. However, they perform poorly in low-lighted backgrounds or when detecting moving objects. To address these issues, we improved the YOLOv5 method and integrated it with other approaches to develop an efficient and accurate system to track the body movement trajectory of macaques to accommodate the fast-moving nature of the animals and maintain high accuracy in low-illumination environments.

2 Proposed Methodology

We first describe the scenarios where our method is applied, then explain how to improve YOLOv5 [15], and finally explain how to process and analyze the collected data. A flowchart was created to present the whole process (Figure 1).

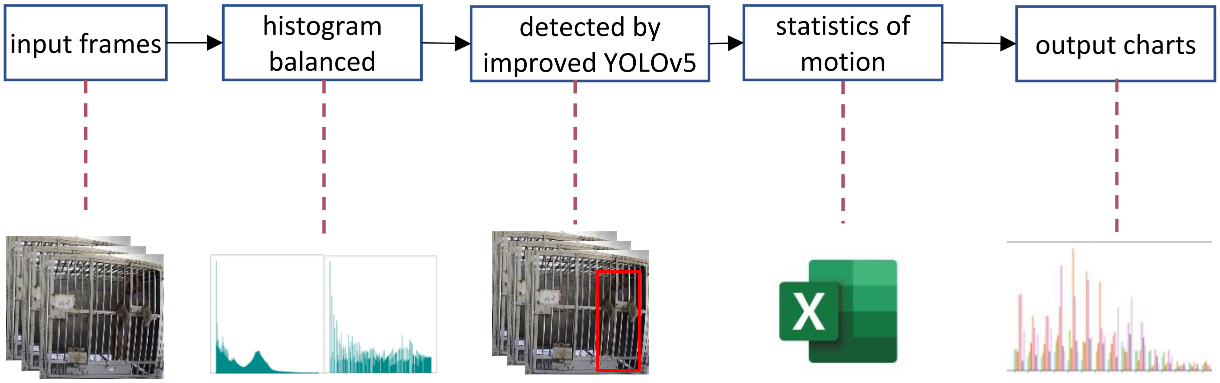

Figure 1. The entire pipeline of the proposed method. First, we converted the video into consecutive video frames and performed histogram equalization for each frame, which is very effective for scenes where it is too bright or too dark. Then, we fed the pre-processed image into the enhanced YOLOv5 model to detect the animal in 2-dimensional space (as shown in the rectangular boxes). We calculated the distance between the rectangular boxes in two consecutive frames and saved it to an excel file. In this way, we could obtain the distance that the animal moved. Finally, we collected all the data for examining the animal's motion preferences and behavior models.

2.1 Pre-Processing with histogram equalization

For each input video, the algorithm needs to manually select a region of interest in the first video frame, such as the entire cage housing the macaque, for further analysis. After getting all the parameters, the algorithm will automatically crop the video size and set the fps. This step removes extraneous factors in video frames to facilitate downstream processing, avoid noise from neighboring cages, and reduce unnecessary computational costs.

Histogram equalization is a simple and effective image enhancement technique that changes the gray level of pixels in the image by changing the histogram distribution of the image. It is usually used to enhance the contrast of images with a small dynamic range. The grayscale distribution of the original image is generally concentrated in a narrow range, so the image contrast is not high enough, which is not conducive to further object detection. For example, if the foreground and background of the image are too dark. Histogram equalization can change the histogram of the original image into a uniform distribution and increase the dynamic range of the gray value difference between pixels to enhance the overall contrast of the picture.

The video recording time in the dataset was distributed over a 24-hour day, including various lighting conditions. Histogram equalization can handle this well. After equalization, foregrounds and backgrounds can be distinguished. as shown in Figure 2.

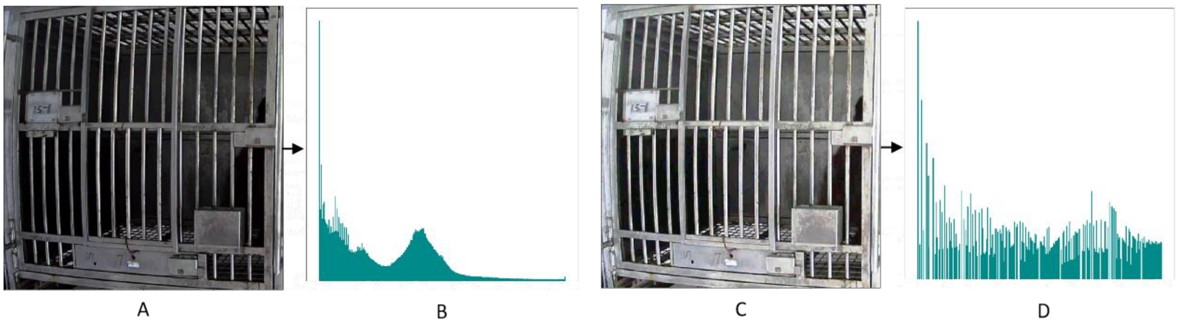

Figure 2. Histogram equalization. (A) The original image. (B) The histogram distribution of A. (C) the image after the histogram equalization of A. The contrast and brightness of the image are significantly improved, especially for scenes that are too bright or too dark. (D) The histogram distribution of A after histogram equalization.

2.2 Improved YOLOv5 Network Structure

General object detection algorithms do not perform well on videos because original object detection algorithms only focus on a single image, and videos contain more temporal and spatial information than simple images. In the target detection of video, the image detection method is directly used to detect the content of the video frame by frame, ignoring the spatiotemporal information and slowing down the detection speed. Each video contains a large amount of redundant data and complementary information between adjacent frames. How to utilize the spatiotemporal context information provided by videos is our concern.

In this paper, we used the improved YOLOv5 [15] model trained in the experiment, and the Temporal Shift Module (TSM) [17] was introduced to extract the temporal information. Because the temporal information was integrated to ensure the quality of spatial features, the features extracted from the network were more robust, and the accuracy could be improved. In addition, in the post-processing stage, we introduced the Kalman filter and Hungarian algorithm, used IOU to measure the similarity of the prediction frame of adjacent frames, optimized the detection results, and then improved the algorithm's robustness to different input videos. The network structure is shown in Figure 3.

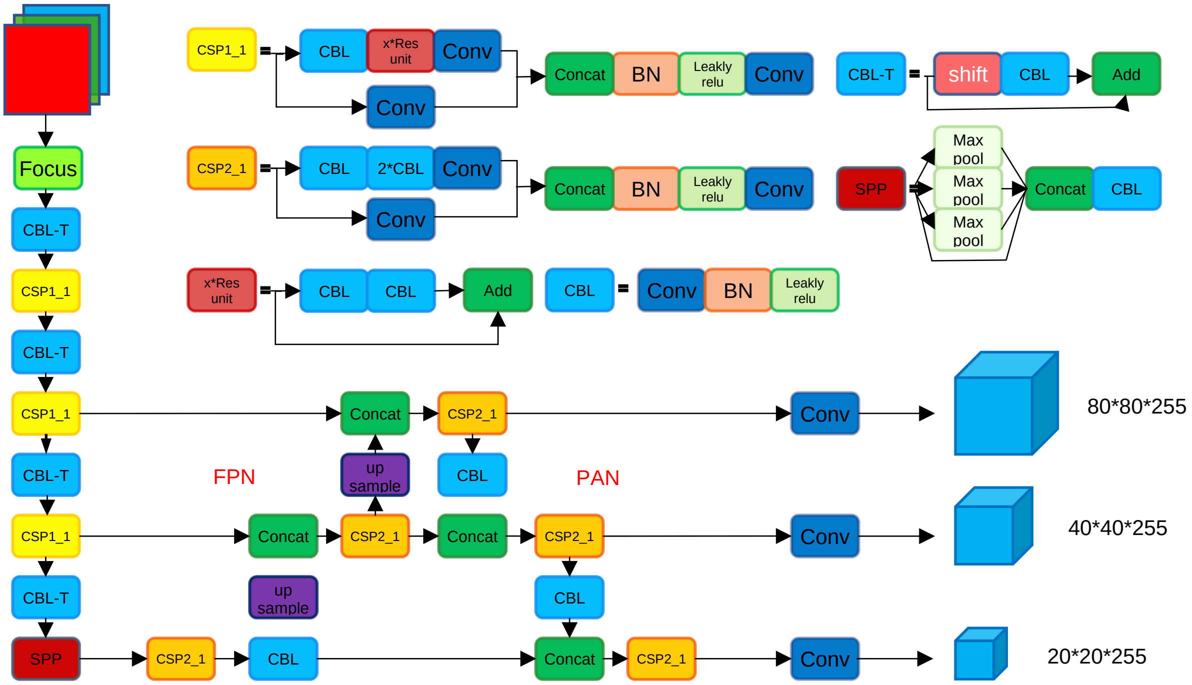

Figure 3. The improved YOLOv5 network

We integrated TSM [17] into the backbone. To reduce the impact on the model's spatial feature learning ability and obtain better temporal features, the backbone performed a shift operation before layers P2, P3, P4, and P5. The specific procedure was to improve the CBL structure in the network and integrate shift modules to CBL in a residual manner. This step exchanged partial channel information between adjacent frames in the time sequence. Experiments demonstrated that this channel information exchange could provide models with temporal modeling capabilities at a lower cost. Therefore, we could obtain high-quality feature maps with temporal-spatial information and improve prediction accuracy. In addition, TSM adopted the mode of residual connection integration, which can reduce the influence of new branches on extracting spatial features.

In this method, the new module (referred to as CBL-T) integrated with the TSM [17] was used to replace the original CBL. This method utilized temporal shifts in the second, third, fourth, and fifth convolution modules of the backbone network and set the parameter batch of the module to 4. The working process of TSM is shown in Figure 4.

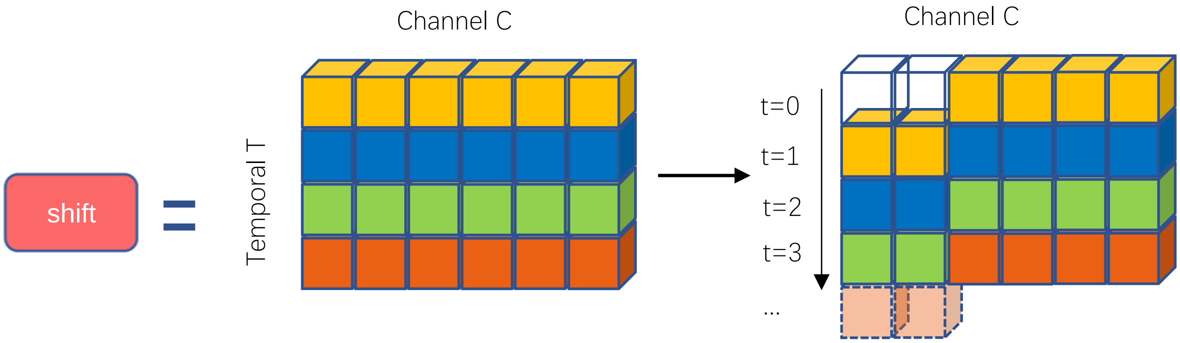

Figure 4. The work process of the Temporal Shift Module

In Figure 4, the same color represents a feature map, the horizontal represents the number of channels, and the vertical represents the time. After moving down one grid, part of the channel information in the yellow feature map will be fused into the blue feature map. It exchanged channel information in the time dimension between adjacent frames and completed time modeling in a low-cost, zero-computation manner. In shift, the parameter of the shift channel slice was set to 1/4. To keep the size of the feature map unchanged, the convolution kernel worked with the size of 1*1 in CBL-T, and the stride was 1.

2.3 Post-Process with Kalman Filter

After the improved YOLOv5 [15] detection was completed, some prediction errors and missed detections might still exist. We employed the Hungarian algorithm and Kalman filter [20] to process the detection results, aiming to remove errors and missing detections. In this approach, the Kalman filter was divided into two phases: 1) the prediction phase and 2) the update phase. In the prediction phase, the Kalman filter listed the detection results of the previous frame as the tracking sequences, named trajectory, and used trajectory as the prior knowledge to predict the status of trajectory in the current frame. Then, we compared the prediction result of the filter with the detection result of the current frame from YOLO using the IOU value. The IOU value measured the degree of overlap between the bounding box generated by individual methods (r_t) and that of the ground truth (r_o) and is defined as:

$$IOU = \frac{|r_t \cap r_o|}{|r_t \cup r_o|} \tag{1}$$

This method requires computing and updating two states of the trajectory: mean and covariance. The mean value represents the target's location information, consisting of the center coordinates of the predicted box, its width and height, and their respective velocity changes. It is represented as $x = [cx, cy, r, h, vx, vy, vr, vh]$.

The covariance represents the uncertainty of the target position information, characterized by an 8x8 diagonal matrix. The larger the number of matrices, the larger the uncertainty. Then, the prediction stage can be expressed as equation 1:

$$x' = Fx \tag{2}$$

In Equation 2, x is the state of track at the time $t-1$ and F is the state transition matrix.

In the update phase, each track that is successfully matched is updated with the corresponding detection in the matched pair. Unmatched tracks and detections may be the error of the filter prediction or the error of the detector itself, and this method will not further distinguish. Unmatched predictions are discarded without any change. The process of the update phase can be expressed as follows:

$$y = z - Hx' \tag{3}$$

In Equation 3, z represents the average value of the detection, that is, the position status information $z = [cx, cy, r, h]$, excluding the speed change value. H is a policy matrix that maps the mean vector of trajectories x' to the detection space. Equation 3 calculates the mean error for detection and tracking.

3 Experiments and Results

3.1 Dataset

Using the ILSVRC2015 [19] dataset format as a standard, we collected 26 video clips of macaques' daily life in a single cage. The length of each video might vary from 20 seconds to 1 minute. Based on the activity characteristics of macaques and to reduce the cost of labeling, we performed the labeling every five video frames. A total of 8000 images were obtained for training. To compare the detection results among different algorithms, we chose another 8000 frames (from 6 video clips) for testing.

3.2 Implementation Details

During training, the initial learning rates were 0.01 and the final learning rates were 0.2. Fifty epochs were performed on one GPU with 16 batches, following [15]. In the training and inference phases, images were resized to a size of 640 pixels on the short side for the feature network. Experiments were performed on a workstation with Intel(R) Core(TM) i7-11700K @ 3.60GHz and NVIDIA GeForce RTX 3090 GPU.

3.3 Comparison Results

We compared our proposed method with several classic algorithms. Except for the mAP score, we defined the success rate as the ratio of frames with an IOU larger than a given threshold in the predicted box to the ground-truth box output by the algorithm. The experimental verification results are shown in Table 1. Our improved algorithm is 8.9% mAP higher than baseline YOLOv5 [15], which shows that our improved strategy is effective, and the success rate is better than that Faster R-CNN [20] and RetinaNet [21]. This means that there are more predictions are closer to the actual value, which shows the superiority of our improved method in video detection. We compared the detection results with and without the Kalman filter [18]. As shown in Figure 5, the green boxes show the real values as a baseline, and the red boxes represent the predicted results without the Kalman filter. When the light is poor and the background is dim, the accuracy of the red boxes decreases significantly. The blue boxes indicate the results with the Kalman filter. We found that the previous problems have been solved better.

Table 1. The comparison between the present method and others.

Methods	mAP (%)	Success (%)
YOLOv5[15]	74.2	79.0
Faster R-CNN[20]	75.6	68.1
RetinaNet[21]	82.6	73.8
Ours	83.1	78.0

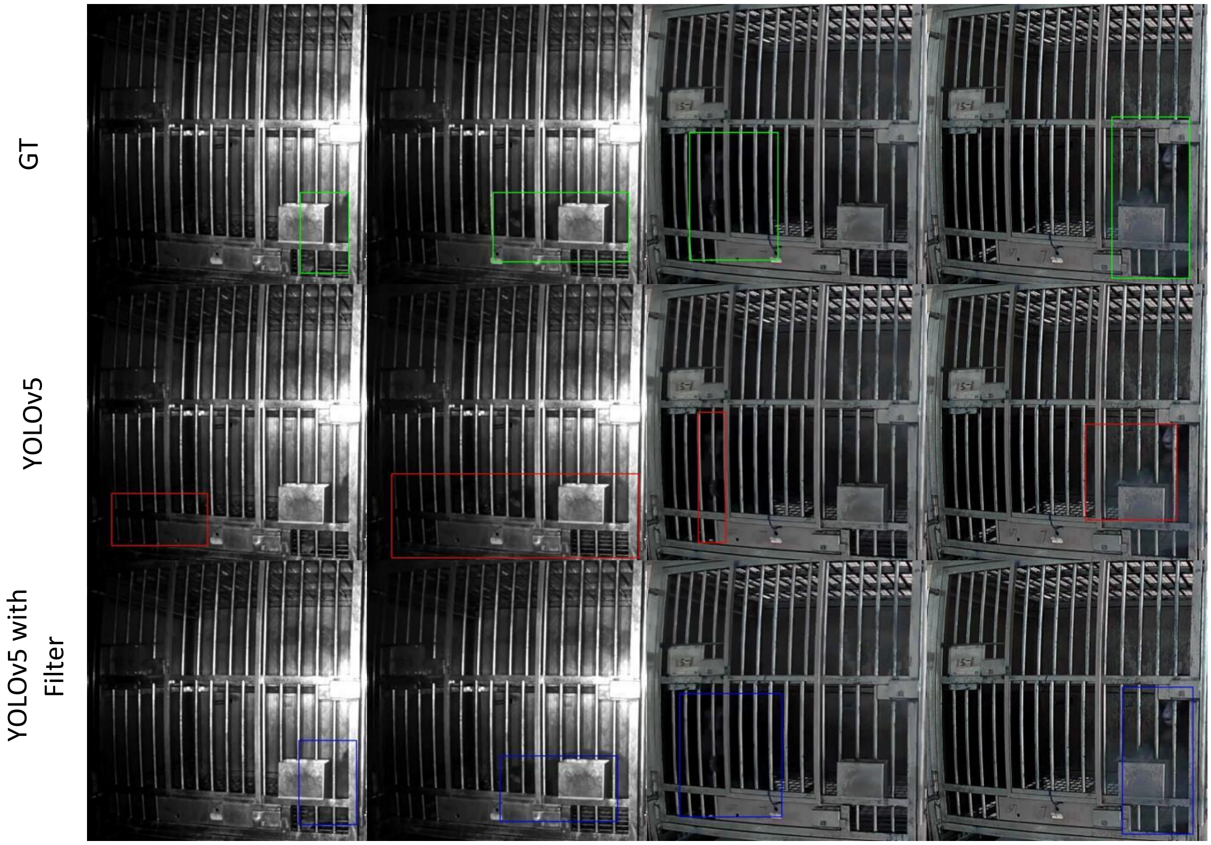

Figure 5. Comparison results of filtering. The top panels show the ground truth value (green rectangles). The middle panels show the output value of YOLOv5 in red. The bottom panels show the output value of YOLOv5 after adding a filter (blue rectangles).

3.4 Ablation experiment

We tested the effectiveness of adding the TSM [17] and Kalman filter [18] into the YOLOv5 [15] model. The experimental comparison data are shown in Table 2. In the improved YOLOv5 network, the introduction of the TSM improved the utilization of timing information by the algorithm, and the mAP increased by about 8%. The Kalman filter was used in the post-processing section to predict and update the detection results, which could deal with detection failures caused by localization errors. The mAP was improved by about 6.5% as compared with the baseline. Also, histogram equalization increased mAP by 5.8% to 80.7%. Finally, by introducing TSM and Kalman Filter into the YOLO algorithm, the final detection result mAP index was about 8.9% higher than the baseline. The TSM and Kalman filter introduced in the present method improved the quality and robustness of feature maps extracted from network models and success rate decreased a little.

Table 2. Results of the baseline method alone and in combination with techniques we used in the present paper. HE: Histogram Equalization, TSM: Temporal Shift Modulle, KF: Kalman Filter.

Methods	mAP (%)	Success (%)
baseline	74.2	79.0
+ HE	80.0	76.4
+ TSM	82.2	78.8
+ KF	80.7	75.7
Ours	83.1	78.0

We compared our method with the baseline based on YOLOv5 alone and its variants (Table 3). Method A is the result of the YOLOv5 alone. It is close to 74.2% mAP. Method B performed the histogram equalization on each input image and was a degenerated variant of the final proposed method. No temporal shift was used in Method B. Method C added the

Temporal Shift Module (TSM) [17] into B. As shown in Table 3, the introduction of TSM improved mAP slightly. Method D is the complete method we proposed. Method E used multi-scale images in the interference and increased mAP by 2.1% to 85.1%. However, the time cost also rose in Method E.

Table 3. Results of baseline method before and after combination with the techniques we used in papers.

Methods	(a)	(b)	(c)	(d)	(e)
histogram equalization		✓	✓	✓	✓
temporal shift			✓	✓	✓
Kalman filtering				✓	✓
multi-scale interference					✓
mAP(%)	74.2	80.0	82.2	83.1	85.1

4 Application

To demonstrate the practical value of our approach in analyzing the behaviors of macaques, we used the method developed in the present study to extract monkeys' movement patterns in daily life cages. We used the trajectories recorded by our method to calculate the activity of two groups of macaques over five weeks. Data for one day at each week interval were collected, including 5x24 hours videos. The method proposed in this paper was applied to these data, and the daytime and nighttime tracking results are shown in Figure 6. Rectangular boxes and lines in Figure 6 represented bounding boxes and trajectories, respectively. The sequence of frames was from left to right, and the time interval between each frame was > 10 seconds. These examples included different activities and different light conditions.

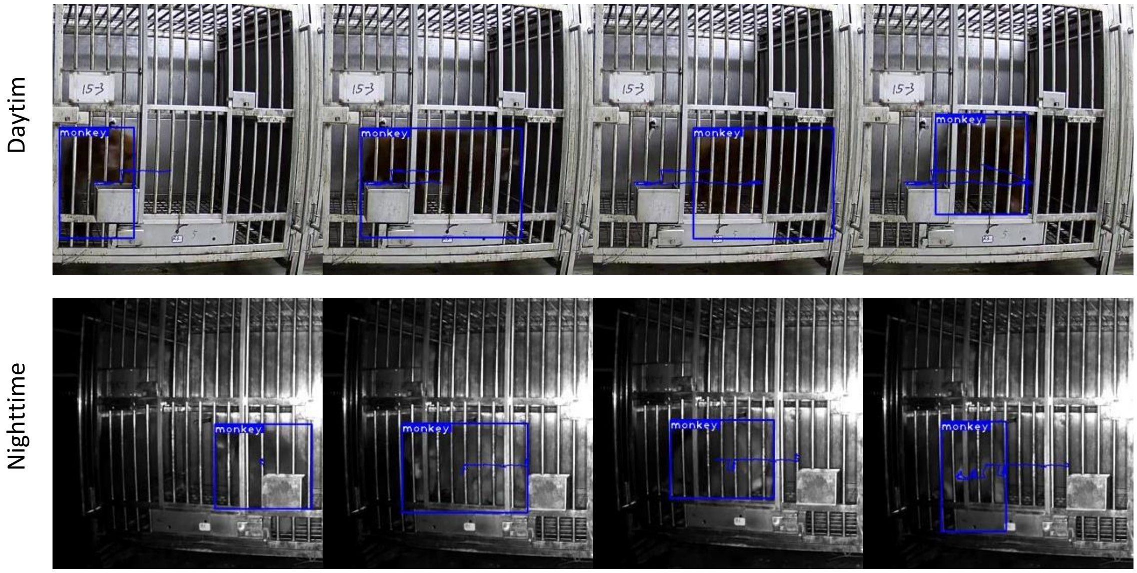

Figure 6. Examples of the daytime and nighttime trajectories

The amount of activities and spatial preference are useful indicators of changes in behavior caused by external conditions such as drug injection and food restriction. Monitoring these parameters can reveal their acute or chronic effects on animals' behaviors. As shown in Figure 7, the distinct activity pattern of sleep-wake is displayed.

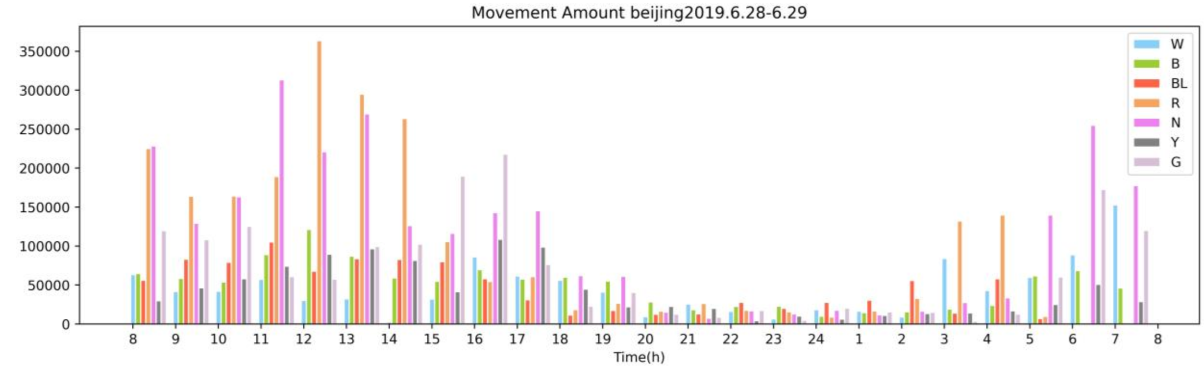

Figure 7. Example of the activity levels of macaques throughout one day

Figure 7 shows the statistical results of seven macaques' activities in one day. We used letters (e.g., W, B, BL, etc.) and colors to distinguish macaques. From the picture, we could see the behavior pattern of macaques from awake to sleep and then awake within 24 hours, and the results were verified to be reliable. Therefore, our method could better detect and track the position of macaques. Then, we calculated the amount of movement and analyzed their behavior patterns. In the future, this method can also be extended to recognition and analyses of macaques' behavior and activity patterns.

5 Conclusion

This paper proposed an improved YOLOv5-based macaque video detection algorithm. Since the video frame sequence is different from the general image sequence, the algorithm introduces sequence convolution to expand the sensitive area, enhance the time modeling ability of the model, and use Kalman filtering to optimize the detection results and the overall network output. The mAP index of this method reached more than 83% on the test set, and the detection accuracy and robustness were good. We also validated the effectiveness of the method in terms of monkey locomotion and applied it to the analysis of real behavioral data. From a practical application point of view, our method can be effectively and accurately used for the analysis of animal position detection box locomotion in a laboratory setting. In the future, the method can be further improved to perform pose estimation and behavior recognition, thus providing more comprehensive information about animal motion patterns.

Acknowledgments

The project was supported by Chinese Academy of Science (QYZDB-SSW-SMC033, XDB32020207), Natural Science Foundation of Anhui Higher Education Institution of China (No. KJ2020A0033).

References

[1] Caiola M, Pittard D, Wichmann T, et al. Quantification of movement in normal and parkinsonian macaques using video analysis[J]. Journal of neuroscience methods, 2019, 322: 96-102. doi: 10.1016/j.jneumeth.2019.05.001.

[2] Yabumoto T, Yoshida F, Miyauchi H, et al. MarmoDetector: A novel 3D automated system for the quantitative assessment of marmoset behavior[J]. Journal of neuroscience methods, 2019, 322: 23-33. doi: 10.1016/j.jneumeth.2019.03.016.

[3] Togasaki D M, Hsu A, Samant M, et al. The Webcam system: a simple, automated, computer-based video system for quantitative measurement of movement in nonhuman primates[J]. Journal of neuroscience methods, 2005, 145(1-2): 159-166. doi: 10.1016/j.jneumeth.2004.12.010.

[4] Hashimoto T, Izawa Y, Yokoyama H, et al. A new video/computer method to measure the amount of overall movement in experimental animals (two-dimensional object-difference method) [J]. Journal of neuroscience methods, 1999, 91(1-2): 115-122. doi: 10.1016/S0165-0270(99)00082-5.

[5] Bala P C, Eisenreich B R, Yoo S B M, et al. Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio[J]. Nature communications, 2020, 11(1): 1-12. doi: 10.1038/s41467-020-18441-5.

[6] Francisco F A, Nührenberg P, Jordan A L. A low-cost, open-source framework for tracking and behavioral analysis of animals in aquatic ecosystems[J]. bioRxiv, 2019: 571232. doi: 10.1101/571232.

[7] Graving J M, Chae D, Naik H, et al. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning[J]. Elife, 2019, 8: e47994. doi: 10.7554/eLife.47994.

[8] Lind N M, Vinther M, Hemmingsen R P, et al. Validation of a digital video tracking system for recording pig locomotor behaviour[J]. Journal of neuroscience methods, 2005, 143(2): 123-132. doi: 10.1016/j.jneumeth.2004.09.019.

[9] Mathis A, Schneider S, Lauer J, et al. A primer on motion capture with deep learning: principles, pitfalls, and perspectives[J]. Neuron, 2020, 108(1): 44-65. doi: 10.1016/j.neuron.2020.09.017.

[10] Ueno M, Hayashi H, Kabata R, et al. Automatically detecting and tracking free-ranging Japanese macaques in video recordings with deep learning and particle filters[J]. Ethology, 2019, 125(5): 332-340. doi: 10.1111/eth.12851.

[11] Walton A, Branham A, Gash D M, et al. Automated video analysis of age-related motor deficits in monkeys using EthoVision[J]. Neurobiology of Aging, 2006, 27(10): 1477-1483. doi: 10.1016/j.neurobiolaging.2005.08.003.

[12] Ballesta S, Reymond G, Pozzobon M, et al. A real-time 3D video tracking system for monitoring primate groups[J]. Journal of neuroscience methods, 2014, 234: 147-152. doi: 10.1016/j.jneumeth.2014.05.022.

[13] Hu G, Cui B, Yu S. Joint learning in the spatio-temporal and frequency domains for skeleton-based action recognition[J]. IEEE Transactions on Multimedia, 2019, 22(9): 2207-2220. doi: 10.1109/TMM.2019.2953325.

[14] Johansson G. Visual perception of biological motion and a model for its analysis[J]. Perception & psychophysics, 1973, 14(2): 201-211. doi: 10.3758/BF03212378.

[15] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing,tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and FranciscoIngham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly, and YouTube integrations, Apr. 2021.

[16] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37. doi: 10.1007/978-3-319-46448-0_2.

[17] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7083-7093. doi: 10.1109/iccv.2019.00718.

[18] Kalman R E. A new approach to linear filtering and prediction problems[J]. 1960. doi: 10.1115/1.3662552.

[19] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.doi: 10.1007/s11263-015-0816-y.

[20] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28. doi: 10.1109/tpami.2016.2577031.

[21] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988. doi: 10.1109/iccv.2017.324.