An Arena-Aware Motion-Based Deep Learning Framework for Automated Rodent Behavior Analysis in the Open Field Test

Ronak Ashvinbhai Sheladiya[1*] and Vishvajit Bakarola[2]

[1] Asha M. Tarsadia Institute of Computer Science and Technology, Uka Tarsadia University, Gujarat, India
ronaksheladiya652@gmail.com* | vishvajit.bakrola@utu.ac.in

Abstract. Automated analysis of rodent behavior in the Open Field Test (OFT) plays a central role in objective and scalable behavioral assessment within neuroscience research. This paper presents a fully automated, video-based framework that performs arena-aware localization, motion feature extraction, and temporal behavior classification directly from raw top-view video recordings. A YOLOv8-based object detection model localizes both the experimental arena and the animal in each frame, enabling arena-relative spatial normalization of motion trajectories. Compact three-dimensional motion feature vectors comprising normalized position and instantaneous speed are grouped into fixed-length temporal sequences and modeled using a Long Short-Term Memory (LSTM) network, yielding frame-aligned behavioral state predictions with associated confidence scores. The proposed pipeline generates ethograms, behavior transition matrices, bout duration statistics, and annotated videos without any reliance on pose estimation or manual annotations. Experimental results demonstrate temporally coherent behavior predictions and stable confidence distributions concentrated above 0.80, confirming the robustness and reproducibility of the approach for automated behavioral analysis across six ethological categories.

Keywords: Open Field Test, Rodent Behavior Analysis, Deep Learning, Motion-Based Features, LSTM, YOLOv8, Computer Vision.

1 Introduction

The Open Field Test (OFT) is among the most widely used paradigms in neuroscience and pharmacological research for characterizing locomotor activity, anxiety, and exploratory behavior in rodents [1][2][3]. Conventional scoring through manual observation or rule-based protocols is time-consuming, subjective, and prone to inter-observer variability, limiting reproducibility and scalability [4][5]. Advances in deep learning have enabled automated, video-based behavioral analysis that delivers objective, high-temporal-resolution measurements with minimal human intervention [6][7][8].

Existing pipelines commonly depend on pose estimation frameworks such as DeepLabCut [7] and SLEAP [22], requiring extensive annotation and degrading under

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_82

occlusion and arena variation [13]. Commercial tools such as Ethovision XT employ background subtraction heuristics that are sensitive to illumination changes [27]. These shortcomings motivate a motion-centric framework that infers behavior directly from video without pose intermediaries [5]. The proposed system combines YOLOv8-based dynamic arena detection with LSTM classification over compact centroid motion features, requiring no annotation and no laboratory-specific calibration. Key contributions are: (i) a YOLOv8 arena-aware detection module; (ii) arena-relative spatial normalization; (iii) a three-dimensional motion feature vector; (iv) an LSTM six-class temporal classifier with per-frame confidence; and (v) a fully automated output pipeline.

2 Related Work

Manual OFT scoring relies on observer-recorded metrics such as distance traveled, center occupancy, and rearing frequency. Although well established, manual methods are susceptible to fatigue, inter-rater disagreement, and limited temporal resolution, particularly in high-throughput experiments [1][4][20]. These limitations have sustained interest in automated video-based analysis [5][6].

Deep learning has advanced automated animal behavior analysis substantially. YOLO-family detectors offer real-time, high-accuracy localization across diverse laboratory conditions [10][11][12]. Pose estimation frameworks including DeepLabCut [7] and SLEAP [22] enable fine-grained body-part tracking but require substantial annotated data and are vulnerable to occlusion. Systems such as JAABA [18] and SimBA [25] combine pose features with supervised classifiers, while MoSeq [8] models behavior as discrete syllables through autoregressive methods. All of these depend on either pose estimation or annotated training data, constraining cross-laboratory generalization. LSTM networks [15], widely adopted for sequential behavior modeling [16][17], combined with spatial detection features consistently outperform frame-wise classifiers [8][18]. The proposed framework avoids pose estimation entirely, operating on centroid motion features to achieve six-class OFT classification without annotation overhead.

3 Technical Approach

The pipeline processes raw OFT video through four sequential phases (Fig. 1): motion feature extraction, temporal sequence construction, LSTM classification, and analysis and visualization.

3.1 Arena and Animal Localization

A custom-trained YOLOv8 model detects both the experimental arena and the rodent in each frame (Fig. 2), selected for its superior inference speed and accuracy under variable illumination [12]. For input frame It, bounding boxes for the arena and animal are defined as:

$$B = \{(x_1, y_1, x_2, y_2)\} \tag{1}$$

where (x1, y1) and (x2, y2) are the top-left and bottom-right corners of the bounding box respectively.

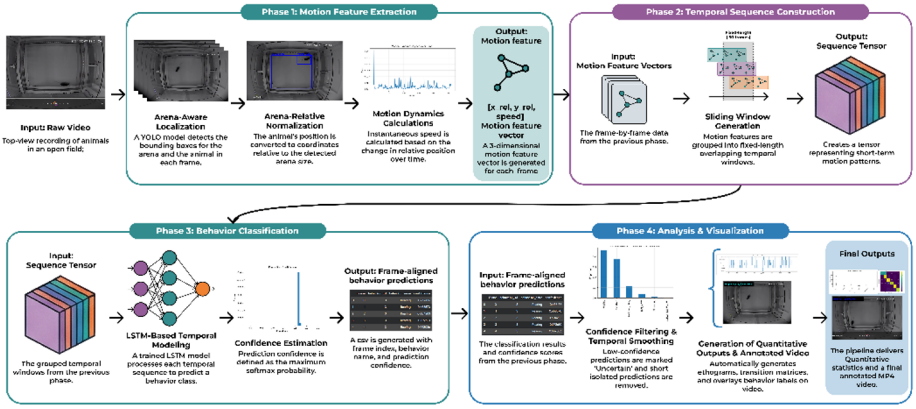

Fig. 1. Overview of the proposed video-based behavioral analysis pipeline.

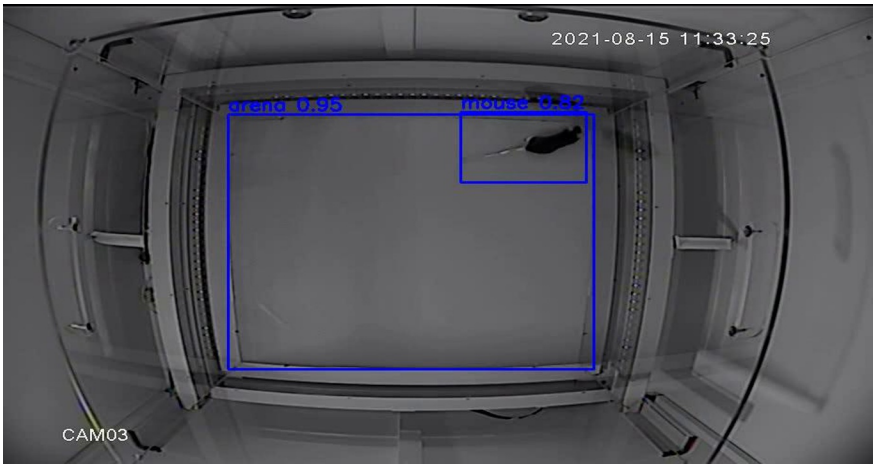

Fig. 2. Arena and mouse localization using the YOLOv8 detector.

3.2 Arena-Relative Normalization, Motion Features, and Temporal Sequences

The animal position is normalized relative to the detected arena to remove camera-induced spatial bias [4,8]. Arena and animal centers are computed as:

$$C^x = \frac{x_1 + x_2}{2}, C^y = \frac{y_1 + y_2}{2} \quad (2)$$

yielding arena-relative normalized coordinates:

$$x_t = \frac{C_m^x - C_a^x}{W_a}, y_t = \frac{C_m^y - C_a^y}{H_a} \quad (3)$$

where Wa and Ha denote the arena bounding box width and height. Resulting trajectories are shown in Fig. 3. Instantaneous displacement is then computed as:

$$\Delta x_t = x_t - x_{t-1}, \Delta y_t = y_t - y_{t-1} \quad (4)$$

and the motion magnitude (speed) is defined as:

$$v_t = \sqrt{((\Delta x_t)^2 + (\Delta y_t)^2)} \quad (5)$$

For each frame, a compact three-dimensional motion-aware feature vector is formed (Figs. 4-5):

$$f_t = [x_t, y_t, v_t] \quad (6)$$

This low-dimensional representation captures both spatial position and movement intensity. Features are grouped using a 30-frame sliding window into temporal sequence tensor:

$$X_t = [f_{t-T+1}, \ldots, f_t] \quad (7)$$

producing a tensor of shape (N, 30, 3), where N is the total number of valid sequences.

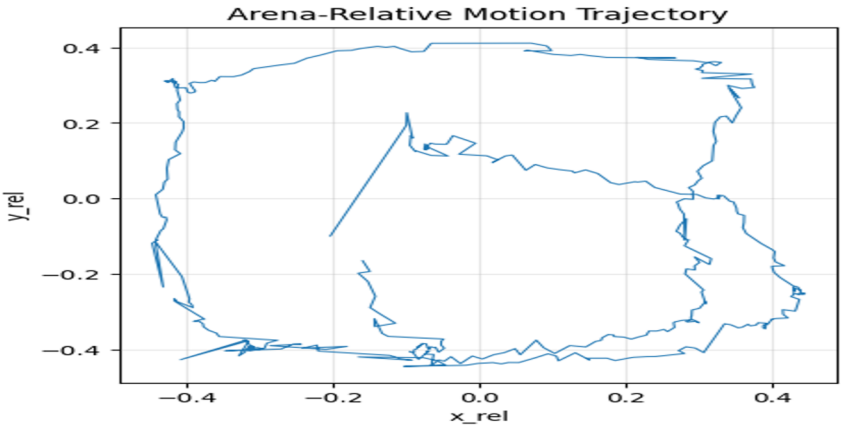

Fig. 3. Arena-relative motion trajectory of the animal after spatial normalization.

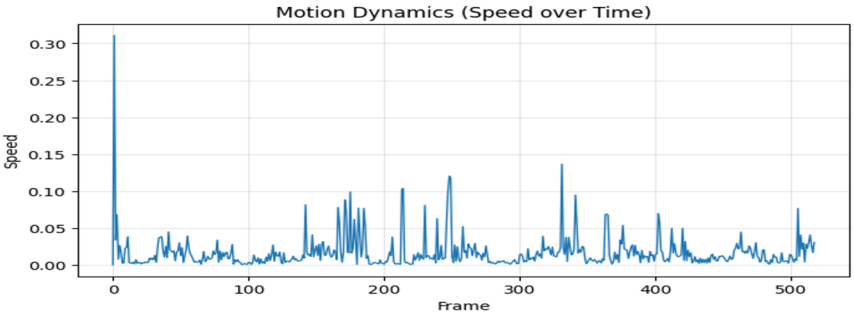

Fig. 4. Instantaneous motion speed computed from normalized centroid displacement over time.

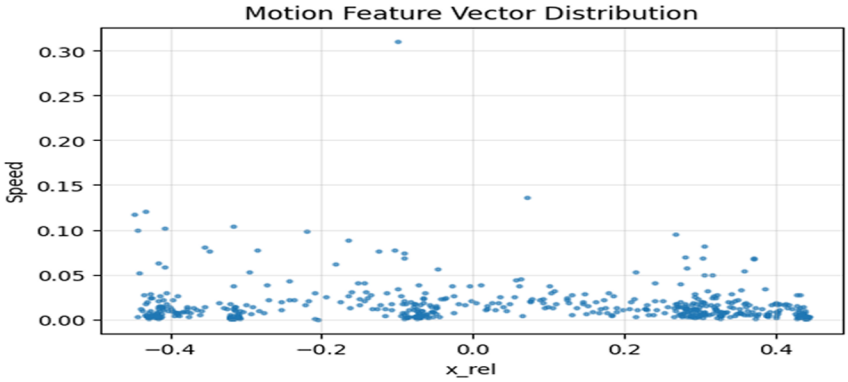

Fig. 5. Distribution of arena-relative motion feature vectors across all analyzed frames.

3.3 LSTM Classification and Confidence

An LSTM network models temporal dependencies within each motion sequence [15]. The hidden state update is:

$$h_t = LSTM(X_t) \tag{8}$$

The final hidden state is projected through a fully connected softmax layer over six predefined behavioral categories:

$$\hat{y}_t = softmax(W h_t + b) \tag{9}$$

where y-hat_t represents the probability distribution over behavioral classes. Prediction confidence is defined as the maximum softmax probability [23]:

$$c_t = max(\hat{y}_t) \tag{10}$$

mes falling below a confidence threshold are flagged as uncertain and excluded from statistical summaries. Final outputs include per-frame labels, ethograms, transition matrices, bout duration statistics, and annotated video overlays.

4 Experimental Setup

Experiments used the publicly available Zenodo OFT dataset [19] comprising top-view recordings of mice under acute and chronic stress paradigms. Only raw RGB video was used; no ground-truth annotations, pose keypoints, or dataset labels were incorporated, enabling unbiased evaluation of motion-based inference [5][6]. Frames were processed without background subtraction or manual cropping [4], with dynamic per-frame YOLOv8 localization handling camera shifts and arena misalignment. Evaluation assessed temporal consistency, behavioral plausibility, and statistical structure through ethograms, transition matrices, and bout duration analysis [8][18], as frame-level ground-truth was unavailable. A fixed window T=30 frames was used with a two-layer LSTM and six-class softmax output, applied uniformly across all recordings.

5 Results and Analysis

The pipeline processes raw OFT video through four sequential phases (Fig. 1): motion feature extraction, temporal sequence construction, LSTM classification, and analysis and visualization.

5.1 Detection and Motion Characterization

YOLOv8 achieved consistent localization with bounding box confidence of 0.95 (arena) and 0.82 (animal) per Fig. 2, adapting dynamically to minor camera shifts. Arena-relative trajectories (Fig. 3) revealed characteristic thigmotaxis with the animal predominantly in peripheral regions - an anxiety marker in OFT [1][3]. Speed profiles (Fig. 4) showed alternating high-speed locomotion and stationary pauses [8], and the feature vector distribution (Fig. 5) confirmed that peripheral positions correspond to lower speeds while central excursions coincide with transient high-speed events.

5.2 Behavioral Classification, Confidence and Transitions

The LSTM predicted frame-level states across six categories: Rearing, Grooming, Periphery_Stay, Fast_Move, Center_Exploration, and Slow_Move. The ethogram (Fig. 6) shows temporally coherent sequences spanning over 14,000 frames, with Periphery_Stay and Rearing as dominant classes consistent with stress-model OFT behavior [2][8]. Prediction confidence (Fig. 7) was sharply concentrated at 0.80-0.85, with approximately 14,000 frames exceeding 0.80. The transition matrix (Fig. 8) revealed strong diagonal dominance (self-transition probabilities 0.88-0.98), confirming temporally stable behavioral predictions with biologically meaningful sparse off-diagonal transitions [8,18].

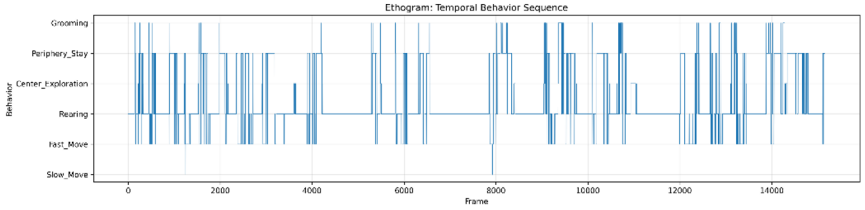

Fig. 6. Ethogram illustrating the temporal evolution of predicted behavioral states across the full recording.

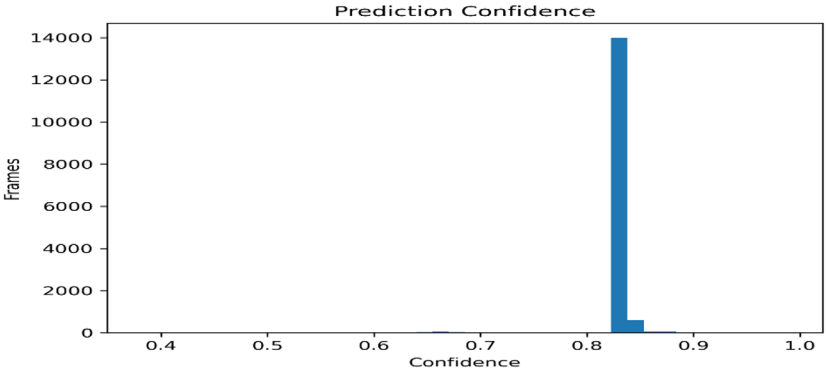

Fig. 7. Distribution of prediction confidence scores across all analyzed frames.

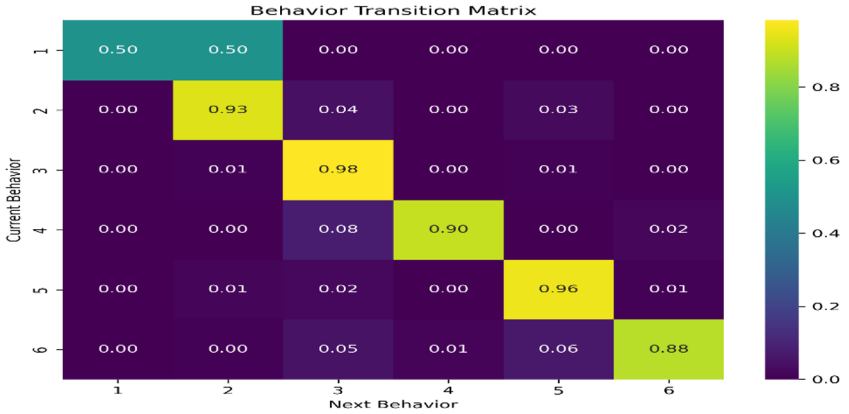

Fig. 8. Normalized transition matrix showing probabilities of behavior-to-behavior Transitions

Fig. 9. Example frame from the automatically annotated Open Field Test video, showing predicted behavioral label and confidence score overlaid on the detected animal in real time.

5.3 Comparison with Existing Methods

Table 1 summarizes a methodological comparison with representative systems. The proposed framework uniquely combines zero annotation requirement, dynamic arena localization, per-frame confidence output, and LSTM temporal classification - advantages not simultaneously present in any compared system. Quantitative benchmarking against manually scored labels remains a valuable direction for future work.

Table 1. Comparison of automated rodent behavior analysis systems.

Method	Pose Est.	Annotation	Arena-Adaptive	Confidence	Classifier
DeepLabCut [7]	Yes	Extensive	No	No	None
SLEAP [22]	Yes	Extensive	No	No	None
JAABA [18]	Yes	Required	No	No	Boosted
SimBA [25]	Yes	Required	No	No	RF/SVM
Ethovision XT [27]	No	Partial	Partial	No	Rule-based
Proposed	No	None	Yes (dynamic)	Per-frame	LSTM (6-class)

6 Discussion

The results confirm that reliable OFT behavioral analysis is achievable from compact centroid motion features without pose estimation. Strong transition matrix diagonal dominance and confidence above 0.80 demonstrate that the LSTM captures genuine temporal behavioral structure - consistent with Wiltschko et al. [8] and Anderson and Perona [5]. Dynamic arena-relative normalization provides cross-laboratory transferability, eliminating recalibration required by Ethovision XT [27] and JAABA [18]. Unlike MoSeq [8] and SimBA [25], which require pose-derived inputs, the supervised LSTM affords direct interpretability and explicit confidence scoring valuable in pharmacological research. The primary limitation is discriminating fine-grained postural behaviors sharing similar speed and position profiles, such as grooming versus freezing; incorporating bounding box aspect ratio changes or optical flow features could address this.

7 Conclusion

This paper presented a fully automated, pose-free framework for OFT rodent behavior analysis using YOLOv8 arena-aware detection and LSTM temporal motion modeling. The system classifies six behavioral categories from compact three-dimensional motion features without annotation or laboratory-specific calibration. Evaluation demonstrated self-transition probabilities of 0.88-0.98, confidence consistently above 0.80, and biologically consistent behavioral patterns - alongside favorable comparison with DeepLabCut, SLEAP, JAABA, SimBA, MoSeq, and Ethovision XT. Future directions include multi-animal tracking, optical flow feature fusion for improved postural discrimination, ground-truth quantitative benchmarking, and attention-based temporal architectures.

Acknowledgments. The authors thank the contributors of the Zenodo OFT dataset [19] for making raw video recordings publicly available for research.

References

1. Walsh, R.N., Cummins, R.A.: The open-field test: a critical review. Psychological Bulletin 83(3), 482-504 (1976)
2. Gould, T.D., Dao, D.T., Kovacsics, J.: The open field test. In: Mood and Anxiety Related Phenotypes in Mice. Humana Press (2009)
3. Seibenhener, M.L., Wooten, M.C.: Open field maze for locomotor and anxiety-like behavior in mice. J. Visualized Experiments 96, 52434 (2015)
4. Spruijt, B.M., DeVisser, A.B., DeVries, P.A.: Pitfalls and principles of behavioral analysis. Neuroscience & Biobehavioral Reviews 38, 1-11 (2014)
5. Anderson, D.J., Perona, P.: Toward a science of computational ethology. Neuron 84(1), 18-31 (2014)

6. Brown, A.E., de Bivort, B.: Ethology as a physical science. Nature Physics 9, 575-580 (2013)
7. Mathis, A., et al.: DeepLabCut: markerless pose estimation with deep learning. Nature Neuroscience 21(9), 1281-1289 (2018)
8. Wiltschko, A.B., et al.: Mapping sub-second structure in mouse behavior. Neuron 88(6), 1121-1135 (2015)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, 436-444 (2015)
10. Redmon, J., et al.: You only look once: unified real-time object detection. In: Proc. IEEE CVPR, pp. 779-788 (2016)
11. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934 (2020)
12. Jocher, G., et al.: Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics (2023)
13. Graving, J.M., et al.: DeepPoseKit: fast and robust animal pose estimation. eLife 8, e47994 (2019)
14. Stephens, G.J., et al.: Dimensionality and dynamics in C. elegans behavior. PLoS Computational Biology 4(4), e1000028 (2008)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735-1780 (1997)
16. Ordonez, F.J., Roggen, D.: Deep LSTM networks for wearable activity recognition. Sensors 16(1), 115 (2016)
17. Donnat, C., et al.: Deep behavioral phenotyping. In: Proc. IEEE ICCV, pp. 895-903 (2018)
18. Kabra, M., et al.: JAABA: interactive machine learning for behavior annotation. Nature Methods 10(1), 64-67 (2013)
19. Sturman, O., et al.: Raw video OFT dataset - acute and chronic stress models. Zenodo. https://doi.org/10.5281/zenodo.8186065 (2023)
20. Crabbe, J.C., Wahlsten, D., Dudek, B.C.: Genetics of mouse behavior. Science 284(5420), 1670-1672 (1999)
21. Delcourt, J., et al.: Video tracking of animal behavior: a review. Behavior Research Methods 45(2), 375-386 (2013)
22. Pereira, T.D., et al.: SLEAP: deep learning for multi-animal pose tracking. Nature Methods 19(4), 486-495 (2022)
23. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. ICML (2017)
24. Sun, J., et al.: Task programming: data efficient behavior representations. In: Proc. IEEE CVPR, pp. 2876-2885 (2021)
25. Nilsson, S.R., et al.: SimBA: open source toolkit for social behavior classification. bioRxiv. https://doi.org/10.1101/2020.04.19.049452 (2020)
26. Shi, F., et al.: Skeleton-based action recognition with directed graph networks. In: Proc. IEEE CVPR, pp. 7912-7921 (2019)
27. Noldus Information Technology: EthoVision XT video tracking software. https://www.noldus.com/ethovision-xt (2023)
28. Dunn, D.J., et al.: Geometric deep learning for 3D kinematic profiling. Nature Methods 18(5), 564-573 (2021)