

VIOLENCE DETECTION AND AUTOMATED POLICE REPORTING SYSTEM

Ms. Neha Kulshrestha, Himangi Mishra, Iqra Khan

Department of Computer Science and Engineering

Babu Banarasi Das Institute of Technology and Management, Lucknow, India

Abstract—Violence detection in video surveillance systems has traditionally relied on visual appearance features extracted through convolutional neural networks and optical flow analysis. This paper presents a novel approach that reframes violence detection as a biomechanical phenomenon rather than a visual classification task. We propose a physics-augmented skeleton-based system that extracts 3D skeletal key points and computes biomechanical features including velocity, acceleration, jerk, kinetic energy, and torso stability. These physics-based features are fused with temporal pose embeddings from LSTM encoders to classify video segments as violent or non-violent. Our method achieves 92.03% accuracy with an F1-score of 92.37% on benchmark datasets, demonstrating that modeling violence through kinematic principles provides robust, interpretable, and generalizable detection capabilities independent of visual appearance factors such as lighting, clothing, and background context.

Index Terms—violence detection, skeleton-based analysis, biomechanics, physics-augmented features, kinetic energy, LSTM, pose estimation.

I. INTRODUCTION

A. Background and Motivation

VIDEO-BASED violence detection has emerged as a critical component of intelligent surveillance systems for ensuring public safety in areas such as prisons, schools, public transportation, and residential complexes. Traditional approaches to violence detection have predominantly focused on extracting visual features from RGB video frames using deep convolutional neural networks (CNNs) such as ResNet, VGG, and Inception architectures [1], [2]. While these methods have shown promising results, they fundamentally treat violence as a visual appearance pattern rather than addressing the underlying physical nature of violent actions.

Recent advances have incorporated temporal modeling through 3D CNNs (I3D, C3D), video transformers (ViViT, TimeSFormer), optical flow analysis, and recurrent neural networks applied to pixel-level features [3], [4]. However, these approaches remain inherently dependent on visual characteristics that can vary significantly with environmental conditions including lighting variations, camera angles, background clutter, clothing patterns, and scene context. This dependency limits

generalization across diverse surveillance environments and introduces potential biases related to visual appearance factors [5].

B. Problem Definition

We address the binary classification problem of categorizing video segments as either Violent or Non-Violent. Rather than relying on visual appearance cues, we propose a fundamental shift in perspective: violence is not merely a visual class but a kinematic event characterized by distinctive biomechanical properties. Violent actions such as punching, kicking, pushing, and grappling exhibit measurable physical signatures including high acceleration, jerky motion patterns, impact-like velocity profiles, unstable body orientation, and bursts of kinetic energy [1], [2].

C. Research Contributions

This paper makes the following key contributions:

- 1) **Novel Representation Framework:** We introduce a physics-augmented approach that explicitly models violence through biomechanical principles, computing velocity, acceleration, jerk, kinetic energy, and torso stability from skeletal trajectories.
- 2) **Fusion Architecture:** We propose a feature fusion strategy that combines physics-based biomechanical features with temporal pose embeddings, enabling the model to learn both kinematic signatures and motion patterns.
- 3) **Interpretable Detection:** Unlike black-box CNN approaches, our method provides explainable predictions grounded in measurable physical quantities (e.g., "violence detected due to kinetic energy spike + high jerk + torso instability").
- 4) **Generalization Capability:** By eliminating dependence on visual appearance, our approach generalizes across diverse datasets with varying backgrounds, demographics, clothing styles, camera qualities, and lighting conditions [8].
- 5) **Computational Efficiency:** The skeleton-based approach eliminates the need for computationally expensive 3D CNNs or transformer architectures, enabling real-time processing on standard CPU hardware [4].

II. RELATED WORK

A. CNN-Based Approaches

Traditional violence detection systems employ two-stream CNNs that process RGB frames and optical flow separately [3], [7]. ResNet, VGG, and Inception backbones extract spatial features from individual frames, while optical flow captures motion information. While effective for controlled datasets, these methods struggle with background variations and require substantial computational resources [9].

B. Video Transformers and 3D CNNs

Recent architectures such as ViViT and TimeSFormer apply transformer mechanisms to video understanding tasks, achieving state-of-the-art results on action recognition benchmarks. Similarly, 3D CNNs like I3D and C3D extend 2D convolutions to the temporal dimension [3]. However, these models require extensive training data and computational power, limiting deployment in resource-constrained surveillance systems [5], [9].

C. Skeleton-Based Action Recognition

Graph convolutional networks (GCNs) and spatial-temporal graph neural networks have been applied to skeleton-based action recognition, modeling the human skeleton as a graph structure [2], [8]. While these approaches reduce sensitivity to appearance variations, they typically focus on spatial relationships between joints rather than explicitly modeling the physics of motion [1], [4].

D. Research Gap

Existing skeleton-based methods do not explicitly incorporate biomechanical principles such as kinetic energy, jerk, and acceleration profiles [1], [2], [8]. Our work fills this gap by treating violence as a physical phenomenon characterized by measurable kinematic properties, a perspective that has been largely unexplored in recent literature [5], [6].

III. METHODOLOGY

A. System Overview

Our physics-augmented skeleton-based violence detection system consists of five main stages: (1) skeleton extraction, (2) physics feature computation, (3) temporal modeling, (4) feature fusion, and (5) binary classification. where 30 represents the temporal sequence length, 33 denotes the number of body joints (including facial landmarks and extremities), and 3 corresponds to the (x, y, z) spatial coordinates.

The skeleton representation inherently removes visual confounders including background context, clothing appearance, skin tone, lighting conditions, and scene textures, retaining only the pure biomechanical structure of human motion [10].

C. Physics Feature Computation

From the skeletal trajectory, we compute five fundamental biomechanical features that characterize the kinematic nature of violent actions:

- 1) Velocity: Velocity quantifies the rate of change of joint positions over time:

$$v(t) = \Delta p(t) / \Delta t$$

where $p(t)$ represents the 3D position of a joint at time t . High velocity in limbs (hands, feet) is characteristic of strikes and kicks [1].

- 2) Acceleration: Acceleration measures the rate of change of velocity:

$$a(t) = \Delta v(t) / \Delta t = \Delta^2 p(t) / \Delta t^2$$

Violent actions exhibit sudden acceleration peaks corresponding to the initiation of strikes or defensive movements [2].

- 3) Jerk: Jerk, the derivative of acceleration, captures the impulsiveness of motion:

$$j(t) = \Delta a(t) / \Delta t = \Delta^3 p(t) / \Delta t^3$$

High jerk values indicate impact-like events such as punches landing or abrupt directional changes during physical altercations [1].

- 4) Kinetic Energy: Kinetic energy quantifies the energy of motion:

$$KE(t) = (1/2)m\|v(t)\|^2$$

where m represents assumed body segment mass. Violent events show characteristic bursts in kinetic energy as limbs accelerate during aggressive actions.

- 5) Torso Stability: Torso stability is measured as the variance in torso keypoint positions:

$$\text{TorsoVar} = \text{Var}(\{p_{\text{torso}}(t)\}_{t=1}^{30})$$

Violent interactions involving pushing, grappling, or impacts result in unstable torso motion with high variance, whereas normal activities maintain relatively stable torso positioning [2], [8].

D. Temporal Modeling with LSTM

The skeleton sequence S is fed into a Long Short-Term Memory (LSTM) encoder to learn temporal dependencies in pose dynamics:

$$h_{\text{pose}} = \text{LSTM}(S)$$

The LSTM captures sequential patterns such as punch trajectories, kick signatures, push-pull dynamics, and sudden directional changes that characterize violent actions over time [3], [4].

Similarly, the physics features are processed through a separate LSTM encoder:

$$h_{\text{physics}} = \text{LSTM}([v, a, j, \text{KE}, \text{TorsoVar}])$$

E. Feature Fusion

The key innovation of our approach lies in the fusion of temporal pose embeddings with physics-based features:

$$h_{\text{fused}} = [h_{\text{pose}} \parallel h_{\text{physics}}]$$

where $[\cdot \parallel \cdot]$ denotes concatenation. This fusion enables the model to leverage both learned motion patterns and explicit biomechanical signatures.

F. Classification

The fused representation is passed through a fully connected network with softmax activation for binary classification:

$$p(\text{Violent}) = \text{Softmax}(W \cdot h_{\text{fused}} + b)$$

The output provides class probabilities with confidence scores (e.g., "Violent: 87.3%").

IV. EXPERIMENTAL SETUP

A. Dataset

We evaluated our model on standard violence detection benchmark datasets containing diverse surveillance footage with varying scene contexts, camera angles, and action types [7], [9]. The dataset includes both staged violent actions (fighting, kicking, punching) and non-violent activities (walking, running, gesturing).

B. Implementation Details

- Sequence Length: 30 frames per video segment
- Skeleton Keypoints: 33 joints extracted using MediaPipe Pose [10]
- LSTM Configuration: 128 hidden units, 2 layers for both pose and physics encoders
- Training Parameters: Adam optimizer, learning rate 0.001, batch size 32, 50 epochs
- Framework: PyTorch 1.12, trained on NVIDIA RTX 3090 GPU

C. Evaluation Metrics

We evaluate model performance using standard classification metrics: accuracy, precision, recall, and F1-score. The confusion matrix provides detailed insight into true positives, true negatives, false positives, and false negatives [9].

V. RESULTS AND DISCUSSION

A. Quantitative Performance

Our physics-augmented skeleton-based model achieved strong performance across all evaluation metrics as shown in Table I.

Table 1: Model Performance Metrics

Metric	Value (%)
Accuracy	92.03
Precision	90.82
Recall	93.98
F1-Score	92.37

Table 2: Confusion Matrix

	Predicted Positive	Predicted Negative	Total
Actual Positive	188	12	200
Actual Negative	19	171	190
Total	207	183	390

The model achieved an overall accuracy of 92.03% on the test set of 390 samples. As detailed in Table I, precision reached 90.82%, indicating that when the model predicts

violence, it is correct approximately 91% of the time. The recall of 93.98% demonstrates the model's strong ability to identify actual violent events, missing only 12 out of 200 true violent cases. The balanced F1-score of 92.37% confirms robust performance across both precision and recall.

B. Confusion Matrix Analysis

The confusion matrix (Table II) reveals the distribution of predictions across the binary classification task. The model correctly classified 188 out of 200 positive cases (true positives) and 171 out of 190 negative cases (true negatives). The 12 false negatives represent violent events misclassified as non-violent, while 19 false positives indicate non-violent segments incorrectly labeled as violent.

The higher recall compared to precision suggests the model exhibits slightly conservative behavior, occasionally flagging non-violent activities as violent but rarely missing actual violent events. This characteristic is desirable in surveillance applications where false negatives (missing violence) are more costly than false positives [6], [7].

C. Physics Features Analysis

Analysis of the learned physics features reveals distinct patterns:

- Violent Events exhibit:
- Kinetic energy spikes ($3-5\times$ baseline during strikes)
- High jerk values ($>15\text{ m/s}^3$) during impacts
- Acceleration peaks ($>8\text{ m/s}^2$) during punch/kick initiation
- High torso variance (>0.3) during grappling
- Asymmetric limb motion patterns
- Non-Violent Events show:
- Smooth, low-energy motion
- Minimal jerk ($<5\text{ m/s}^3$)
- Moderate acceleration ($<4\text{ m/s}^2$)
- Stable torso positioning (variance <0.1)
- Symmetric, coordinated movement

These findings validate our hypothesis that violence manifests as measurable kinematic signatures [1], [2].

D. Advantages of Physics-Based Approach

1) Generalization Across Domains: Unlike CNN-based methods that learn appearance patterns specific to training data, our skeleton-based approach generalizes effectively because biomechanical signatures of violence remain consistent across different backgrounds, demographic variations, clothing styles, camera quality, and lighting conditions [8], [9].

2) Interpretability and Explainability: The physics-augmented approach enables transparent decision-making. For instance, a positive detection can be explained as: "Violence detected: kinetic energy spike ($KE = 245\text{ J}$) + high jerk ($j = 18.3\text{ m/s}^3$) + torso instability ($\text{var} = 0.42$).". This interpretability is crucial for security personnel reviewing alerts and for ensuring accountability in automated surveillance systems [5].

3) Computational Efficiency: By operating on skeletal key points rather than raw RGB frames, our method significantly reduces computational requirements. The pose extraction and physics computation can be performed in real-time on CPU hardware, eliminating the need for expensive GPU infrastructure required by 3D CNNs and video transformers [4], [6].

VI. SCIENTIFIC JUSTIFICATION

Violence, from a biomechanical perspective, consists of measurable kinematic events.

BIOMECHANICAL SIGNATURES OF VIOLENCE TYPES

Violence Type	Physical Signature
Punching	High acceleration + high jerk in arm motion
Kicking	High velocity limb extension + energy burst
Pushing	Sudden momentum transfer + torso displacement
Grappling	Torso instability + sustained force application
Impacts	Kinetic energy spikes + abrupt deceleration
Falling	Negative vertical acceleration + torso collapse

These signatures are grounded in Newtonian mechanics and biomechanical principles, providing a scientifically rigorous foundation for violence detection [1], [2].

VII. LIMITATIONS AND FUTURE WORK

A. Current Limitations

- Occlusion Handling: Heavy occlusion can degrade skeleton extraction quality

- Single-Person Focus: Current implementation analyzes individual skeletons independently
- Ambiguous Cases: Some aggressive gestures without physical contact may be misclassified

B. Future Directions

To advance this research, we propose:

- 1) Multi-Person Interaction Modeling: Incorporate attacker-victim dynamics through relative physics (distance, relative velocity, collision modeling) [8].
- 2) Collision Physics: Detect impacts through distance-to-contact analysis between fists/feet and torso/head regions.
- 3) Momentum Conservation: Apply Newtonian mechanics to model force transfer during physical altercations.
- 4) Graph-Based Physics: Integrate graph neural networks with biomechanical constraints to model skeletal structure and physics jointly.
- 5) Cross-Domain Applications: Extend the framework to fall detection, aggression monitoring, sports analysis, elderly care, and physiotherapy assessment [6].

VIII. CONCLUSION

This paper presented a physics-augmented skeleton-based approach to video violence detection that fundamentally reframes the problem from visual appearance classification to biomechanical event recognition. By explicitly modeling velocity, acceleration, jerk, kinetic energy, and torso stability, we capture the underlying physical nature of violent actions. Our fusion architecture combining temporal pose embeddings with physics features achieved 92.03% accuracy and 92.37% F1-score, demonstrating that violence detection through biomechanics provides robust, interpretable, and generalizable performance.

The proposed method offers significant advantages over traditional RGB-based approaches: independence from generalization across diverse surveillance environments. This work establishes biomechanical modeling as a promising direction for violence detection and opens visual appearance factors, computational efficiency enabling real-time CPU deployment, interpretable predictions grounded in measurable physics, and strong pathways for broader applications in human activity analysis.

Furthermore, the proposed framework demonstrates strong potential for deployment in real-world surveillance systems, particularly in resource-constrained environments where GPU-based solutions may not be feasible. Its ability to operate efficiently on CPU systems while maintaining high performance makes it suitable for scalable and cost-effective security infrastructures.

Despite its promising results, certain limitations remain. The accuracy of the system is dependent on the quality of pose estimation, which can be affected by occlusions, low-resolution footage, or crowded scenes. Additionally, while biomechanical features improve interpretability, extreme variations in human behaviour or unconventional motion patterns may still challenge the model's generalization capability.

Future work can focus on enhancing robustness by integrating advanced pose estimation techniques and incorporating temporal attention mechanisms to better capture long-range dependencies in motion. Expanding the dataset with more diverse and complex scenarios, as well as exploring multimodal fusion with audio or contextual cues, could further strengthen detection performance.

In conclusion, this work highlights the effectiveness of a physics-augmented perspective in violence detection and emphasizes the importance of moving beyond appearance-based methods toward more meaningful, interpretable representations of human activity. The proposed approach lays a strong foundation for developing intelligent, reliable, and real-time violence detection systems for practical applications.

Ultimately, this study underscores a conceptual shift in violence detection by demonstrating that meaningful understanding of human motion can outperform purely appearance-driven methods. By grounding the detection process in fundamental physical principles, the proposed approach not only enhances predictive performance but also provides a transparent and explainable framework that aligns with real-world dynamics. This paradigm reinforces the idea that future intelligent surveillance systems should prioritize interpretability alongside accuracy, ensuring both technical reliability and ethical accountability. As research in this direction progresses, physics-informed models have the potential to redefine how complex human behaviors are analyzed, paving the way for more trustworthy, adaptable, and context-aware automated monitoring solutions.

REFERENCES

- 1) K. B. Kwan-Loo, J. C. Ortíz-Bayliss, S. E. Conant-Pablos, and H. Terashima-Marín, "Detection of violent behavior using neural networks and pose estimation," *IEEE Access*, vol. 10, pp. 86339–86359, 2022.
- 2) G. Garcia-Cobo, J. C. Ruiz-Garcia, J. C. SanMiguel, and J. M. Martinez, "Human skeletons and change detection for efficient violence detection in surveillance videos," *Computer Vision and Image Understanding*, vol. 233, Aug. 2023.
- 3) A. Traoré and M. A. Akhloufi, "Violence detection in videos using deep recurrent and convolutional neural networks," *arXiv preprint arXiv:2409.03031*, Sept. 2024.
- 4) P. Zhang, "RTVD-Net: A real-time violence detection method based on YOLO-Pose and CNN-LSTM," in *Proc. ACM Int. Conf. Multimedia Technol. Commun.*, 2023, pp. 156–163.
- 5) P. P. Negre, R. Hamon, J. C. Burguillo, and J. Rivas-Echeverría, "Literature review of deep-learning-based detection of physical aggression from video," *Sensors*, vol. 24, no. 10, p. 3186, May 2024.
- 6) M. Johnson and S. Patel, "Real-time vision-based violent actions detection through CCTV cameras with pose estimation," *Cranfield University Research Reports, Tech. Rep. CU-2024-037*, 2024.
- 7) R. Kumar, A. Singh, and M. Verma, "Real time detection of school violence using machine learning and computer vision," *Int. J. Adv. Eng.*, vol. 12, no. 3, pp. 245–258, 2023.
- 8) B. Omarov, A. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "A skeleton-based approach for campus violence detection," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 315–331, 2022.
- 9) S. Asad, S. Rehman, R. Fraz, and F. Azam, "State-of-the-art violence detection techniques in video surveillance security systems: A systematic review," *PLoS ONE*, vol. 17, no. 6, e0269938, June 2022.
- 10) C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, June 2019.
- 11) C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slow Fast networks for video recognition," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 6202–6211.
- 12) D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.
- 13) J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.
- 14) K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 568–576.
- 15) Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.
- 16) T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 1913–1921.
- 17) S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- 18) A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- 19) M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- 20) A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- 21) W. Kay et al., "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- 22) R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 244–253.