

Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications

{BE Students}* Mhaske Pragati Sambhaji, Patil Sakshi Balraje, Sable Aishwarya
Ramhari, Sawant Dipali Bhagwat
{Faculty}* Tanuja Shrikant Dhage

Department of Computer Engineering, DattaKala Group Of Institution Faculty Of
Engineering, Bhigwan-413130, University of Pune, Maharashtra, INDIA.

pragatimhaske2@gmail.com

dipalibhagwat015@gmail.com

aishwaryasable59@gmail.com

sakshibalrajepatil123@gmail.com

Abstract

Surveillance cameras are increasingly utilized worldwide, but the large volume of generated video data poses challenges for real-time monitoring. This paper presents a deep learning-based approach for automatic violence detection in surveillance footage. By leveraging 3D convolutional neural networks and transfer learning from pre-trained action recognition models, we propose an efficient and accurate system for identifying violent activities. Our experimental evaluation on multiple datasets demonstrates improved classification accuracy with fewer model parameters compared to state-of-the-art methods.

Keywords: Violence Detection, Video Surveillance, Deep Learning, Anomaly Detection, Human Activity Recognition, Security, Smart Cities, 3D Convolutions, Action Recognition.

Introduction

In recent years, the proliferation of surveillance cameras has enhanced public security by monitoring human activities in real-time. Surveillance footage plays a crucial role in crime prevention, law enforcement, and forensic investigations. However, manually analyzing extensive video streams is inefficient and prone to human error. The need for automated violence detection has driven research towards deep learning-based solutions that can efficiently analyse surveillance footage. Traditional methods for violence detection relied on hand-crafted features such as Histogram of Oriented Gradients (HOG) and optical flow descriptors. However, these methods often fail in complex real-world scenarios, such as varying lighting conditions, occlusions, and background clutter. With advancements in deep learning, convolutional neural networks (CNNs) have emerged as powerful tools for video analysis, enabling automatic feature extraction and classification.

Human Activity Recognition (HAR) has gained significant attention in recent years, leveraging sensor data to classify human actions. Early HAR approaches used 2D CNNs, but these methods lacked temporal information critical for recognizing dynamic activities like violence. Recent developments in 3D CNNs have facilitated the extraction of spatiotemporal features, significantly improving recognition accuracy.

Despite progress, challenges remain in deploying violence detection models for real-world applications. High computational costs, model efficiency, and the ability to generalize across diverse datasets are key concerns. This study proposes a computationally efficient deep learning framework that enhances violence detection accuracy while maintaining real-time processing capabilities.

Literature Review

In this paper, we propose a novel method to automatically generate low-level spatio-temporal descriptors showing good performance, for high-level human-action recognition tasks. We address this as an optimization problem using genetic programming (GP), an evolutionary method, which produces the descriptor by combining a set of primitive 3D operators. As far as we are aware, this is the first report of using GP for evolving spatio-temporal descriptors for action recognition. In our evolutionary architecture, the average cross-validation classification error calculated using the support-vector machine (SVM) classifier is used as the GP fitness function. (Li Liu, Ling Shao, Peter Rockett August 2012) [1]

In this survey, we will show some research developed by the academic community and some projects developed for the industry. We intend to show the basic principles to begin developing applications using Kinect, and present some projects developed at the VISGRAF Lab. And finally, we intend to discuss the new possibilities, challenges and trends raised by Kinect. (Leandro Cruz, Djalma Lucio, Luiz Velho August 2012) [2]

We propose a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. We take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Our large and highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. Finally we generate confidence-scored 3D proposals of several body joints by reprojecting the classification result and finding local modes. (Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, Andrew Blake June 2011) [3]

this paper, we propose a new skeletal representation that explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space. Since 3D rigid body motions are members of the special Euclidean group $SE(3)$, the proposed skeletal representation lies in the Lie group $SE(3) \times \dots \times SE(3)$, which is a curved manifold. Using the proposed representation, human actions can be modelled as

curves in this Lie group. Since classification of curves in this Lie group is not an easy task, we map the action curves from the Lie group to its Lie algebra, which is a vector space. (Raviteja Vemulapalli, Felipe Arrate, Rama Chellappa June 2014) [4]

In this paper, we propose a self-supervised contrastive learning method to learn video feature representations. In traditional self-supervised contrastive learning methods, constraints from anchor, positive, and negative data pairs are used to train the model. In such a case, different samplings of the same video are treated as positives, and video clips from different videos are treated as negatives. Because the spatio-temporal information is important for video representation, we set the temporal constraints more strictly by introducing intra-negative samples. (R. Chaudhry, F. Ofili, G. Kurillo, R. Bajcsy, and R. Vidal June 2013) [5]

When combined with several other cost-effective designs including separable spatial/temporal convolution and feature gating, our system results in an effective video classification system that produces very competitive results on several action classification benchmarks (Kinetics, Something-something, UCF101 and HMDB), as well as two action detection (localization) benchmarks (JHMDB and UCF101-24). (J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu June 2019) [6]

We present pure-transformer based models for video classification, drawing upon the recent success of such models in image classification. Our model extracts spatiotemporal tokens from the input video, which are then encoded by a series of transformer layers. In order to handle the long sequences of tokens encountered in video, we propose several, efficient variants of our model which factorise the spatial- and temporal-dimension. (A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid October 2021) [8]

In this paper, we propose Temporal Alignment Module (TAM), a novel few-shot learning framework that can learn to classify a previous unseen video. While most previous works neglect long-term temporal ordering information, our proposed model explicitly leverages the temporal ordering information in video data through temporal alignment. This leads to strong data-efficiency for

few-shot learning. (K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles June 2020) [9]

The novelty is revealed in three aspects. First, it is a unique utilization of particle trajectories for modelling crowded scenes, in which we propose new and efficient representative trajectories for modelling arbitrarily complicated crowd flows.

Gap Analysis

Existing violence detection systems have several limitations that hinder their effectiveness in real-world applications. Many models rely heavily on manually annotated datasets, which can introduce biases and reduce generalizability. Additionally, most current methods require significant computational resources, making real-time deployment impractical for surveillance systems with limited hardware capabilities. The lack of diverse datasets also impacts model robustness, leading to decreased performance when applied to varied environments and camera perspectives. Furthermore, existing models often struggle with video compression artifacts and environmental variations, affecting their accuracy in real-time scenarios. Our approach addresses these challenges by leveraging transfer learning, optimizing 3D CNN architectures, and conducting extensive cross-dataset validation to improve model generalizability and robustness.

Existing System

Current violence detection systems rely on manual monitoring or traditional machine learning methods using handcrafted features like HOG and HOF. While these approaches offer some level of detection, they are limited by their inability to effectively capture complex spatiotemporal patterns in videos. Additionally, traditional systems struggle with varying environmental conditions, occlusions, and real-time processing constraints. Some deep learning models, such as CNNs and LSTMs, have been introduced for violence detection, but they often require high computational power and lack generalization across diverse datasets. These limitations highlight the need for a more efficient and accurate approach to automated violence detection.

Second, chaotic dynamics are introduced into the crowd context to characterize complicated crowd motions by regulating a set of chaotic invariant features, which are reliably computed and used for detecting anomalies. Third, a probabilistic framework for anomaly detection and localization is formulated. (B. Zhao, L. Fei-Fei, and E. P. Xing June 2011) [10]

Proposed System

The proposed system integrates deep learning-based violence detection into automated video surveillance systems, ensuring efficient and accurate classification of violent activities. The system consists of a preprocessing module, where video frames are extracted using uniform temporal sampling, resized, and normalized to maintain consistency. A feature extraction component utilizes a pre-trained 3D CNN model (X3D-M) to capture spatiotemporal patterns associated with violent actions. The classification module refines these extracted features using fully connected layers to determine whether a video segment contains violent activity. Finally, a post-processing step aggregates segment-wise predictions over time, reducing false alarms and improving detection accuracy. By optimizing computational efficiency and leveraging transfer learning, our system is well-suited for real-time surveillance applications.

Methodology

1.Dataset Preparation: Collection and extension of seven publicly available video datasets. Manual annotation of additional datasets to ensure diverse and challenging video samples.

2.Model Selection & Architecture: Use of an optimized 3D Res Net-based CNN (X3D-M) for feature extraction.

Two approaches:

Fine-Tuned X3D-M Model (FT): Trains the entire model, including pre-trained weights.

Transfer-Learned X3D-M Model (TL): Extracts features using a frozen X3D-M model and trains additional fully connected layers.

3.Training & Optimization: Binary classification setup with violence and non-violence labels.

Pre-processing: Frame extraction, normalization, and resizing for input consistency.

Loss function: Binary Cross-Entropy (BCE).

Optimization: Adagrad with an initial learning rate of $1e-3$.

4.Performance Evaluation: Accuracy (ACC) and Area Under Curve (AUC) metrics. Cross-validation tests (one-on-one and leave-one-out) to analyse model generalization.

Evaluation under video compression artifacts to simulate real-world streaming scenarios.

System Architecture

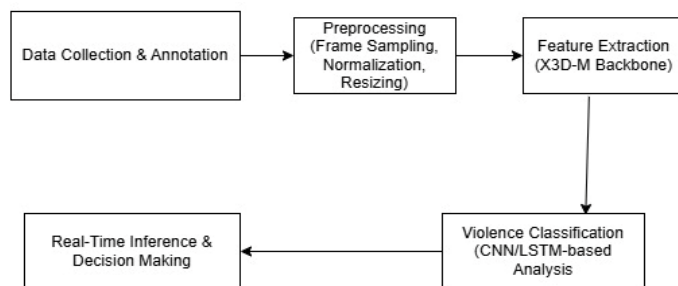
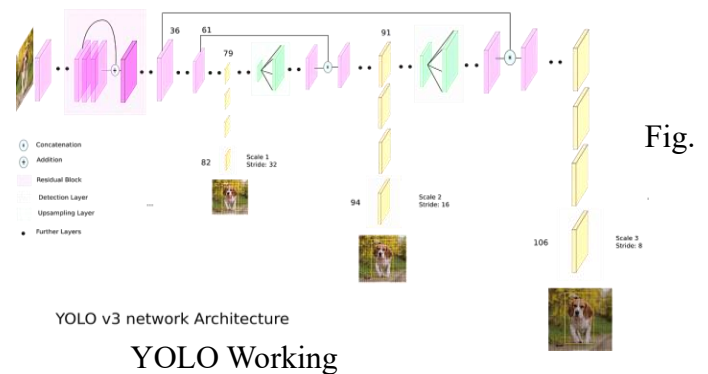


Fig. Violence Detection

Algorithm

- YOLO takes an input image and divides it into a grid.
- The image is divided into a grid of cells, and each cell is responsible for predicting objects located within it.
- For each bounding box, YOLO predicts class probabilities for a fixed number of object classes. These probabilities indicate the likelihood of the detected object belonging to each class.



Future Scope

- 1.Focus on developing deep learning-based models for automated violence detection.
- 2.Use of 3D convolutional neural networks (CNNs) for spatiotemporal video analysis.
- 3.Evaluation on multiple datasets to ensure model robustness.

Conclusion

This study presents a computationally efficient deep learning approach for violence detection in surveillance videos. By leveraging 3D CNNs and transfer learning, our system achieves high accuracy with reduced computational complexity. Experimental results demonstrate superior performance compared to existing methods, making our approach suitable for real-world deployment. Future work will focus on enhancing model generalizability, integrating additional datasets, and optimizing deployment for edge computing environments.

Acknowledgement





I am heartily thankful to my project guide Prof. T. S. Dhage for her valuable guidance and inspiration. In spite of her busy schedule she devoted herself and took keen and personal interest in giving me constant encouragement and timely suggestion. It gives me a great pleasure in presenting the Review Paper of my project. I take opportunity to express my deep sense of gratitude to our Principal Dr. S. A.

Patil who helps us. I would be failing our duty if I do not thank our HOD Dr. S. S. Bere for his word of encouragement and special guidance and vital inspiration.

References

1. L. Liu, L. Shao, and P. Rockett, "Genetic programming-evolved spatiotemporal descriptor for human action recognition," in Proc. Brit. Mach. Vis. Conf., 2012, pp. 1–12.
2. L. Cruz, D. Lucio, and L. Velho, "Kinect and RGBD images: Challenges and applications," in Proc. 25th SIBGRAPI Conf. Graph., Patterns Images Tuts., Aug. 2012, pp. 36–49.
3. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in Proc. CVPR, Jun. 2011, pp. 1297–1304.
4. R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in Proc. IEEE Conf. Compute. Vis. Pattern Recognition., Jun. 2014, pp. 588–595.
5. R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bioinspired dynamic 3D discriminative skeletal features for human action recognition," in Proc. IEEE Conf. Compute. Vis. Pattern Recognition. Workshops, Jun. 2013, pp. 471–478.
6. J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu, "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics," in Proc. IEEE/CVF Conf. Compute. Vis. Pattern Recognition. (CVPR), Jun. 2019, pp. 4006–4015.
7. S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in Proc. Eur. Conf. Compute. Vis. (ECCV), 2018, pp. 305–321.
8. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 6836–6846.
9. K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in Proc. IEEE/CVF Conf. Compute. Vis. Pattern Recognition. (CVPR), Jun. 2020, pp. 10618–10627.
10. B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in Proc. CVPR, Jun. 2011, pp. 3313–3320.

Author Bibliography

	Mhaske Pragati Sambhaji Birth Place: Pune(2002), Currently pursuing B.E. in computer Technology from DGOI FOE. Certification in Python Basic and Advance Technology Dhayri Pune
	Patil Sakshi Balraje Birth Place: Karmala (2003), Currently pursuing B.E. in computer Technology from DGOI FOE. Certification in Web Technology from ProAzure Technology, Kharadi Pune.
	Sable Aishwarya Ramhari Birth Place: Beed(2002), Currently pursuing B.E. in Computer Technology from DGOI FOE. Certification in Robotic Process Automation from ProAzure Technology, Kharadi Pune.
	Sawant Dipali Bhagwat Birth Place: Beed (2004), received diploma from TSSM BSOER Polytechnic Narhe Pune, Currently pursuing B.E.in computer Technology from DGOI FOE. Certification in Python Basic and Advance Technology Dhayri Pune.
	Tanuja Shrikant Dhage Birth place: Ahmedabad (Gujrat) Education: B.E. Computer from JSPM Wagholi(2015)

	<p>M.E. Computer from Dattakala Group of Institutions Faculty of Engineering, Bhigwan(2022) Currently working as Assistant Professor (Computer IT) dept in Dattakala Group of Institutions Faculty Engineering,Bhigwan.</p>
	<p>Sachin Sukhdev Bere working as Associate Professor in Dattakala Group of Institutions Faculty of Engineering Bhigwan. He completed his Ph.D. in Computer Science and Engineering, from the, Shri Jagdishprasad Jhabarmal Tibrewala University, Rajasthan India. And completed his M.Tech (CSE) with First class & Distinction from JNTU-Hyderabad affiliated college. He has 14 years of teaching experience, 7 years of Research experience. Presently he is working as an Associate Professor in Dattakala Group of Institutions Faculty of Engineering Bhigwan, Maharashtra. He published almost 15 research articles in reputed journals and conferences. His interesting areas are Machine Learning, Artificial Intelligence, Deep Learning Techniques and Programming Languages.</p>