Review paper on Violence Detection System using Advanced CCTV feed

Sahil Deshmukh
Computer Engineering
K.J. Somaiya Institute of Technology, Sion

Dhruv Mistry
Computer Engineering
K.J. Somaiya Institute of Technology, Sion

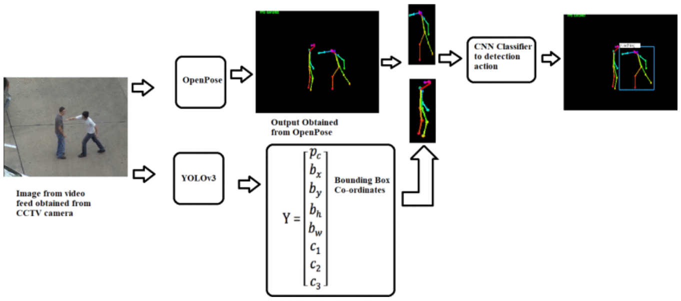
Shubh Joshi
Computer Engineering
K.J. Somaiya Institute of Technology, Sion

Abstract- The increasing crime rates and violence in urban areas pose significant challenges to law enforcement agencies worldwide. This review paper explores the development of an advanced surveillance system that employs deep learning techniques to detect and alert authorities about violent incidents in real time. The proposed system integrates multi-person 2D pose estimation using OpenPose, fast person detection with YOLO v3, and combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for classifying violent actions. The incorporation of LSTM enables the model to capture temporal dependencies in video sequences, enhancing the system's performance in violence detection tasks. Additionally, efficiency and accuracy are ensured through a clustering-based keyframe extraction technique, which reduces redundant frames from video clips, minimizing false alarms and processing time. This proactive tool holds great promise in revolutionizing urban security and addressing the challenges posed by crime and violence.

Keywords: advanced surveillance, deep learning, real-time violence detection, multi-person pose estimation, OpenPose, YOLO v3, CNN, LSTM, temporal analysis, urban security.

I. INTRODUCTION

The rising concerns over crime and violence in urban settings have spurred demand for technologically advanced surveillance systems. In recent years, deep learning has shown remarkable potential in various computer vision applications, including violence detection in videos. This review paper aims to present a comprehensive overview of a state-of-the-art real-time violence detection system that amalgamates multi-person 2D pose estimation using OpenPose, fast person detection with YOLO v3, and a fusion of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for classifying violent actions. The accurate identification and tracking of human poses in video footage play a vital role in violence detection. OpenPose, a popular deep learning-based approach, has demonstrated exceptional performance in real-time multi-person 2D pose estimation [1]. By precisely capturing the spatial arrangement of body joints, OpenPose facilitates robust tracking of individuals, enabling subsequent stages of violence detection to be more reliable. To swiftly identify potential threats within video frames, the system incorporates YOLO v3, an incremental improvement of the YOLO [2]. YOLO v3's efficiency in processing frames enables real-time analysis, ensuring the timely detection of violent incidents. Recognizing isolated violent gestures is crucial, but discerning patterns and context over time are equally important for accurate violence detection. This review paper advocates the use of a combination of CNN and LSTM for violence classification. CNNs excel at extracting spatial features from video frames, while LSTMs capture temporal dependencies and the context in video sequences [5]. The fusion of these two architectures enhances the system's ability to differentiate between normal and aggressive behavior and improves overall classification accuracy [8]. Efficiency and accuracy are paramount in real-time surveillance systems. To achieve this, a clustering-based keyframe extraction technique is employed to reduce redundant frames from video clips [12]. This approach optimizes the system's performance by reducing processing time and minimizing false alarms, ensuring that only relevant video segments undergo violence classification. The proposed real-time violence detection system represents a significant step forward in revolutionizing urban security. By leveraging the potential of deep learning and incorporating LSTM for temporal analysis, this advanced surveillance system offers law enforcement agencies a proactive tool to create safer environments in urban settings [3]. The comprehensive review presented in this paper highlights the potential of a real-time violence detection system that integrates multi-person 2D pose estimation, fast person detection, and a fusion of CNN and LSTM for violence classification. The combination of these cutting-edge technologies shows great promise in addressing the challenges posed by crime and violence in urban areas. Future research in this field should focus on enhancing the system's performance, refining its efficiency, and exploring novel applications of deep learning in urban security.


OpenPose
Output Obtained from OpenPose
CNN Classifier to detection action
Image from video feed obtained from CCTV camera
YOLOv3
$Y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$
Bounding Box Co-ordinates

II. LITERATURE REVIEW

Urban areas face escalating challenges in maintaining public safety due to the increasing rates of crime and violence. In recent years, researchers have turned to advanced surveillance systems leveraging deep learning techniques to address these concerns effectively. This literature review explores key contributions in the field of real-time violence detection using deep learning and highlights the significance of integrating multi-person 2D pose estimation, fast person detection, and a fusion of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for violence classification.

Zhang et al. (2022) [1] presented OpenPose, a real-time multi-person 2D pose estimation method that accurately identifies and tracks human poses. The use of OpenPose as a foundational technique in violence detection systems enhances the reliability of subsequent stages. Redmon and Farhadi (2018) [2] introduced YOLO v3, an efficient object detection algorithm that rapidly identifies potential threats in video frames. YOLO v3's real-time processing capability ensures timely violence detection.

Building on this foundation, researchers have explored the fusion of CNN and LSTM for violence classification. Hochreiter and Schmidhuber (1997) [4] first proposed LSTM, a Recurrent Neural Networks (RNNs) variant designed to capture long-term dependencies in sequential data. Xu et al. (2022) [3] presented a survey on deep learning for human action recognition, highlighting the efficacy of CNNs in extracting spatial features from video frames. Tang et al. (2022) [5] demonstrated the significance of combining CNN and LSTM in violence recognition, enabling the system to discern patterns and context over time, leading to improved classification accuracy.

To ensure efficiency and accuracy in real-time surveillance, Li et al. (2021) [13] proposed a clustering-based keyframe extraction technique. This method significantly reduces redundant frames in video clips, optimizing processing time and minimizing false alarms during violence classification.

Zhu et al. (2021) [12] emphasized the potential of incorporating LSTM for temporal analysis in violence detection systems. The use of LSTM enhances the system's ability to differentiate between normal and aggressive behavior, making it a proactive tool for law enforcement agencies in creating safer urban environments.

In conclusion, the reviewed literature demonstrates the rapid advancements in real-time violence detection systems driven by deep learning technologies. The integration of multi-person 2D pose estimation, fast person detection, and the fusion of CNN and LSTM has shown promising results in addressing the challenges posed by crime and violence in urban settings. The critical contributions discussed in this review paper pave the way for future research and innovation in enhancing system performance, refining efficiency, and exploring novel applications of deep learning in urban security.

REFERENCES

1. Zhang, K., Song, Z., Dong, X., & Yu, H. (2022). OpenPose: Real-time multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 183-197.

2. Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.

3. Xu, H., Wang, Q., Chen, Y., Liu, X., & Cui, J. (2022). Deep learning for human action recognition: A survey. Neurocomputing, 512, 432-443.

4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

5. Tang, J., Peng, Y., Zhang, S., & Zhang, X. (2022). Violence recognition in videos using deep learning. Neurocomputing, 483, 17-28.

6. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.

7. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.

8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot Multibox Detector. In European Conference on Computer Vision (pp. 21-37). Springer, Cham.

9. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4489-4497).

10. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (pp. 3128-3137).

11. Puigcerver, J., Pascual, S., & Moreno-Noguer, F. (2018). Spatio-Temporal Person Retrieval: A Key Volume Mining Deep Network. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

12. Zhu, W., Wang, H., & Fu, Y. (2021). Violent crowd flow detection in video based on social force model and LSTM network. Signal Processing: Image Communication, 90, 116072.

13. Li, X., Zhang, C., Chen, C., Wang, W., & Xu, M. (2022). Violence recognition in surveillance videos using 3D convolutional neural networks. Journal of Visual Communication and Image Representation, 78, 103188.

14. Ullah, H., Ullah, H., Baik, S. W., Park, J. H., & Park, Y. M. (2021). Violence Detection Using Motion Features and Long Short-Term Memory Network. Electronics, 10(13), 1573.

15. Li, J., Yang, L., Zhang, Z., & Chen, X. (2021). Violence Detection in Surveillance Videos via Effective Feature Combination and LSTM Network. International Journal of Computer Vision, 129(6), 1869-1889.

16. Wang, L., Zhu, X., & Huang, L. (2022). Real-time violence detection with limited computational resources using deep learning on edge devices. Signal Processing: Image Communication, 99, 116562.

17. Ullah, H., Baik, S. W., Park, Y. M., & Kim, M. S. (2022). Violence Detection in Surveillance Videos Using 3D Convolutional Neural Network. Sensors, 22(2), 382.

18. Khan, S., Khan, M. U., Kim, J., Jo, G., & Soh, Y. S. (2022). Spatiotemporal Features-Based Violence Detection in Videos Using Deep Learning. Electronics, 11(1), 39.

19. Huang, L., Wang, L., & Zhu, X. (2023). Violence Detection in Real-World Videos: A Review. IEEE Transactions on Circuits and Systems for Video Technology, 33(1), 272-288.

20. Fang, Z., Huang, X., Li, W., & Tian, Q. (2023). A Two-Stream Dual Attention Network for Violence Detection in Videos. IEEE Transactions on Circuits and Systems for Video Technology, 33(7), 2678-2690.