International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 5, Sep.-Oct, 2024 pp: 08-13 ISSN: 2584-2145
www.ijemh.com

Factors Influencing Real-Time Violent Behavior Recognition: A Review

Xiuliang Zhang, Chuntao Liu, Tadiwa Elisha Nyamasvisva
[1]*Student, Infrastructure University Kuala Lumpur, Xiuliang Zhang*
[2]*Student, Infrastructure University Kuala Lumpur, Chuntao Liu*
[3]*Student, Infrastructure University Kuala Lumpur, Tadiwa Elisha Nyamasvisva*
Corresponding Author: Xiuliang Zhang

Date of Submission: 02-09-2024 Date of Acceptance: 14-09-2024

ABSTRACT: Real-time violent behavior recognition is pivotal for enhancing public safety by enabling prompt detection and intervention in potentially dangerous situations. This review comprehensively examines the factors influencing the effectiveness of real-time recognition systems. We discuss the essential role of real-time recognition in improving security measures and highlight critical factors affecting system performance. Key areas covered include the quality and diversity of datasets, which are crucial for training robust models; the design of model architectures, which impacts the balance between computational efficiency and recognition accuracy; and the constraints imposed by computational power, memory, and network bandwidth. By analyzing these factors, this review provides insights into the current challenges and advancements in real-time violent behavior recognition, offering a foundation for future research and development aimed at enhancing system performance and reliability.

KEYWORDS: Real-time, Violent Behavior Recognition，Dataset quality, Model architecture, Computational constraints.

I. INTRODUCTION

Violent behavior recognition refers to the automated detection and identification of aggressive or harmful actions through video or sensor data[1]. This technology is designed to monitor and analyze behaviors in public environments in real-time, allowing for the prompt detection of potential violent events. Violent behavior recognition systems typically rely on computer vision and machine learning algorithms, which extract features from video streams and classify behaviors based on those features[2, 3]. This technology is significant for enhancing public safety, as it enables security personnel to quickly respond to and address potential threats, thereby reducing the occurrence and impact of violent incidents.

Traditional surveillance methods mainly depend on manual monitoring and analysis, which have significant limitations. Firstly, human monitoring requires considerable time and effort, and operators may overlook potential threats due to fatigue. Secondly, the vast amount of surveillance data makes it difficult for manual analysis to cover all video streams, leading to possible missed or false detections. Moreover, manual monitoring lacks real-time responsiveness and is unable to swiftly address sudden violent incidents. These limitations reduce the effectiveness and efficiency of traditional methods in enhancing public safety, especially in large-scale public spaces and complex environments where comprehensive security is difficult to ensure.

With advancements in computer vision and artificial intelligence technologies, modern violent behavior recognition systems have rapidly evolved. These technologies utilize deep learning and big data analysis to automatically extract and analyze behavioral features from video data, enabling efficient and accurate detection of violent behaviors[4-6]. Modern systems not only improve detection accuracy but also enhance real-time processing capabilities, allowing for quicker responses to violent events[7]. In the context of violent behavior detection, real-time recognition is crucial, as it allows the system to swiftly capture potential aggressive or harmful actions and issue alerts almost simultaneously as the behavior occurs. The key role of real-time recognition lies in its ability to significantly shorten the response time between the occurrence of the behavior and the subsequent intervention, which is essential for effectively managing and preventing violent incidents[8]. For example, in public spaces where violent behavior is detected in real time, security personnel can immediately intervene to prevent escalation and protect individuals on-site[5].

Compared to traditional surveillance methods, real-time violent behavior detection

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 5, Sep.-Oct, 2024 pp: 08-13 ISSN: 2584-2145
www.ijemh.com

systems can automatically process and analyze large volumes of video data, quickly detecting abnormal behaviors through advanced algorithms and models[9]. This rapid response capability enables security personnel to receive alerts as soon as the behavior begins, allowing for immediate intervention. This not only improves operational efficiency but also enhances the ability to respond to sudden incidents[10]. Real-time recognition technology reduces reliance on manual monitoring, improves accuracy and reliability, and significantly enhances overall public safety management[11].

Currently, real-time violent behavior recognition technology is rapidly advancing, driven by improvements in deep learning, computer vision, and hardware processing capabilities[12]. However, despite significant technological progress, several challenges remain. First, the lack of diversity and representativeness in existing datasets limits the generalization capability of models. Second, while feature extraction techniques and model architectures have matured, achieving further improvements in real-time processing efficiency while maintaining recognition accuracy remains a challenge[13]. Furthermore, ensuring efficient processing under resource-constrained environments, including limitations in computational power, memory, and network bandwidth, requires further exploration.

The goal of this review is to systematically analyze and summarize the key factors influencing real-time violent behavior recognition, aiding researchers and practitioners in gaining a deeper understanding of the current state and challenges in this field. Through a comprehensive review of existing research, this paper will explore various aspects such as dataset quality, feature extraction techniques, model architecture design, and resource limitations to identify the core factors affecting the performance of real-time violent behavior recognition systems in practical applications. Based on these analyses, this paper aims to provide a reference for future technological improvements and innovations, laying the foundation for the development of more efficient real-time recognition systems.

II. Impact of Dataset Quality on Real-Time Violent Behavior Recognition

In the field of real-time violent behavior recognition, the quality of datasets plays a crucial role in determining the accuracy of the system. A high-quality dataset should possess diversity and representativeness to ensure that models can effectively detect violent behavior across various complex scenarios. However, existing violent behavior recognition datasets vary significantly in terms of diversity, data scale, and labeling quality, all of which directly impact the performance of recognition systems[12].

The diversity and quality of a dataset largely dictate how well a model is trained and its ability to generalize in real-world scenarios. For instance, the Hockey Fight Dataset focuses on detecting violent behavior in ice hockey games, where most of the video clips capture fighting incidents in a sports context. The advantage of this dataset lies in its specific setting and clear representation of violent actions, enabling models to achieve high accuracy in detecting aggression within such environments. However, due to the strong contextual limitation, models trained on this dataset may struggle to generalize to non-sports settings, where violent behavior manifests differently, reducing their effectiveness in broader public safety applications.

In contrast, the Crowd Violence Dataset encompasses a wider range of public space incidents, including riots and clashes, offering more diverse scenarios and more complex forms of violent behavior. This diversity enhances a model's generalization capability, making it better suited for detecting various types of aggression in different environments. However, the complexity and variability of group behaviors also introduce new challenges: the boundaries between violent and non-violent actions in crowd situations can be ambiguous, leading to potential false positives or false negatives. Therefore, while the dataset excels in terms of richness, the precision of labeling and the clarity of behavior definitions remain areas for improvement.

Additionally, the RWF-2000 dataset, which has gained popularity in recent years, includes 2,000 video clips from surveillance cameras, covering both violent and non-violent behavior in public and private settings[14]. A key advantage of this dataset is its collection of real-world surveillance footage, offering high-quality data that closely simulates real-life application environments, thus making model training more applicable to practical needs. However, the relatively small scale of the RWF-2000 dataset means it cannot encompass all possible violent behavior patterns, which may cause models to underperform when faced with unfamiliar scenarios.

Despite the importance of these datasets in supporting violent behavior recognition, they still face several limitations and challenges. First, the

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 5, Sep.-Oct, 2024 pp: 08-13 ISSN: 2584-2145
www.ijemh.com

existing datasets are limited in scale, especially when it comes to handling rare or complex forms of violent behavior. This scarcity of examples can lead to suboptimal model performance. Second, the definitions of violent behavior and the labeling standards across datasets are not uniform. In some cases, violent behavior labels are ambiguous, making it difficult for models to accurately distinguish between violent and non-violent actions during training. Additionally, many datasets are contextually narrow, lacking the diversity and complexity necessary for models to perform well in real-world applications, where environments and behaviors are highly variable.

In summary, the quality and diversity of datasets have a profound impact on the performance of real-time violent behavior recognition systems. Future research should focus on developing more representative and higher-quality datasets to enhance model generalization across different settings, improving the accuracy and reliability of detection in complex, real-world scenarios.

III. Model Architecture Design Impact on Real-Time Processing

In real-time violent behavior recognition, the design of the model architecture is crucial for ensuring both the system's processing capability and overall performance[15]. Real-time recognition demands the system to process large-scale video data within limited time and resources, thus requiring models to balance efficiency and complexity[16]. There are significant differences between lightweight models and complex deep neural networks in terms of parameters, computational requirements, and recognition performance[17]. Effective model design, which balances accuracy with processing speed, is key to achieving efficient real-time recognition.

Lightweight models like MobileNet and EfficientNet are widely used in resource-constrained environments. For example, MobileNetV2 has only about 3.4 million parameters and requires approximately 300 million FLOPs (floating-point operations). This model leverages depthwise separable convolutions to reduce computational load while maintaining high recognition accuracy. Such an architecture is well-suited for real-time violent behavior recognition, especially in scenarios with limited hardware resources, such as embedded devices or edge computing platforms. The strength of lightweight models lies in their ability to enhance processing speed and minimize latency while maintaining sufficient recognition accuracy.

In contrast, complex deep neural networks like ResNet-50 and Inception-v3 excel in recognition accuracy but require significantly more parameters and computational resources. For instance, ResNet-50 has about 25.6 million parameters and requires over 4 billion FLOPs, far more than the MobileNet series. This makes ResNet-50 more suitable for running on high-performance servers or dedicated hardware. However, in real-time scenarios, complex models may introduce processing delays due to their longer computation and inference times, potentially compromising real-time performance.

The complexity of the model architecture directly affects system latency and response times. Models like Tiny-YOLO, which are designed specifically for real-time applications, offer a good balance between efficiency and accuracy. Compared to standard YOLO models, Tiny-YOLO drastically reduces parameters (with 8.7 million parameters in Tiny-YOLOv3, compared to 62 million in YOLOv3), while maintaining reasonable detection accuracy. This enables Tiny-YOLO to respond quickly in real-time scenarios, making it suitable for violent behavior detection and monitoring.

Besides parameter count, the inference time of the model is another critical factor impacting real-time recognition performance. Lightweight models typically have shorter inference times than complex ones. For instance, MobileNetV3 has an inference time of under 20ms on standard mobile devices, whereas complex models like ResNet-50 may exceed 100ms. This time difference is particularly important in real-time settings, where longer inference times may prevent the system from promptly detecting violent behaviors.

To further enhance real-time processing capabilities, several architectural optimization techniques are employed, such as model pruning, quantization, and knowledge distillation. These techniques help reduce model parameters and computational demands without significantly compromising performance. For instance, model pruning eliminates redundant neuron connections to lighten computational load, while quantization reduces model weights from 32-bit floating-point numbers to 8-bit integers, speeding up inference.

In conclusion, the design of the model architecture plays a critical role in real-time violent behavior recognition. Lightweight models such as MobileNet and Tiny-YOLO excel in resource-constrained environments by providing efficient real-time processing. Although complex models can achieve higher recognition accuracy, their

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 5, Sep.-Oct, 2024 pp: 08-13 ISSN: 2584-2145
www.ijemh.com

computational complexity can hinder real-time performance. By balancing accuracy with processing speed and leveraging optimization techniques, real-time violent behavior recognition systems can be better adapted to various application scenarios.

IV. Impact of Resource Constraints on Real-Time Violent Behavior Recognition

Resource limitations, device performance, and computational power are critical factors that directly impact the system's ability to process data efficiently and deliver timely and accurate responses[18]. From a theoretical perspective, these factors influence the speed of data processing, the efficiency of model inference, and the overall response time, determining whether the system can function effectively under constrained resources[19].

Firstly, computational resource limitations are a major challenge for real-time recognition performance. These systems often need to process large volumes of video data with high-frequency feature extraction and model inference[20]. While deep learning models (e.g., Convolutional Neural Networks, Transformers) offer high recognition accuracy, they require significant computational power, especially GPUs or powerful CPUs. On high-performance devices (e.g., servers or dedicated GPU hardware), complex models can run quickly and provide accurate recognition results. For instance, models like ResNet have tens of millions of parameters, yielding high precision but also demanding substantial computational resources. In resource-constrained environments (e.g., mobile devices or embedded systems), inference time increases dramatically, reducing real-time capabilities. Hence, lightweight models such as MobileNet or Tiny-YOLO are often employed to balance computational demands and recognition performance.

Secondly, device performance determines how much data the system can process and at what speed. Device performance is mainly reflected in the processor's computational power, memory capacity, and data transmission capability. In terms of computational power, the floating-point operations per second (FLOPs) of GPUs are a key metric for evaluating their processing capacity. High-end GPUs like the NVIDIA V100 can reach hundreds of teraflops, while embedded devices such as Jetson Nano have only about 500 gigaflops of power. Insufficient computational power leads to increased inference times, making it challenging to meet real-time requirements, particularly when handling high-resolution video or multi-camera data. To achieve efficient processing in resource-limited environments, techniques such as model compression, pruning, quantization, and knowledge distillation are commonly used to reduce model complexity and computational load.

Memory limitations also significantly impact system performance. Real-time violent behavior recognition systems need to handle multiple video streams simultaneously, performing image preprocessing, feature extraction, and model inference, all of which consume substantial memory[21]. If a device's memory capacity is insufficient, the system may be unable to load the complete model or process large-scale data, resulting in delays or crashes during recognition. Traditional deep learning models typically require several gigabytes of memory for training and inference, whereas lightweight models can achieve satisfactory performance with less memory usage. Furthermore, memory constraints can affect the parallelism of model inference, further slowing down overall processing speed.

Network bandwidth plays a crucial role in real-time systems, particularly in distributed or edge computing architectures. When video data needs to be transmitted from endpoint devices to servers for processing, insufficient bandwidth can cause transmission delays, thereby affecting the system's response time. For instance, transmitting high-definition video streams usually requires a high-bandwidth connection, and inadequate bandwidth can prevent the system from receiving video data in real-time, leading to latency or frame drops. To address this issue, many systems offload some processing tasks to edge devices, leveraging edge computing to reduce data transmission demands and optimize real-time performance.

In summary, resource limitations, device performance, and computational power are critical factors in the design and deployment of real-time violent behavior recognition systems. By selecting appropriate model architectures, optimizing hardware configurations, and employing edge computing and model compression techniques, it is possible to build efficient real-time recognition systems in resource-constrained environments. However, the ongoing challenge remains in balancing recognition accuracy with real-time

processing performance, an area that requires continuous research and innovation..

V. CONCLUSION

This review has examined key factors influencing real-time violent behavior recognition, emphasizing its crucial role in enhancing public safety through timely detection and intervention. The accuracy and effectiveness of these systems depend significantly on three primary factors: dataset quality, model architecture, and resource constraints. High-quality, diverse datasets are essential for training models that generalize well across different scenarios, while lightweight model architectures balance computational efficiency and recognition accuracy, critical for real-time performance. Additionally, addressing resource constraints such as computational power, memory, and network bandwidth is vital for optimizing system performance. Future advancements should focus on improving these areas to enhance the responsiveness and effectiveness of real-time violent behavior recognition systems.

REFERENCES

[1]. Bermejo Nievas, E., et al. Violence detection in video using computer vision techniques. in Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14. 2011. Springer.

[2]. Zhou, P., et al., Violence detection in surveillance video using low-level features. PLoS one, 2018. **13**(10): p. e0203668.

[3]. Accattoli, S., et al., Violence detection in videos by combining 3D convolutional neural networks and support vector machines. Applied Artificial Intelligence, 2020. **34**(4): p. 329-344.

[4]. Sernani, P., et al., Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access, 2021. **9**: p. 160580-160595.

[5]. Sumon, S.A., et al., Violence detection by pretrained modules with different deep learning approaches. Vietnam Journal of Computer Science, 2020. **7**(01): p. 19-40.

[6]. Yue-Hei Ng, J., et al. Beyond short snippets: Deep networks for video classification. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[7]. Ullah, F.U.M., et al., A comprehensive review on vision-based violence detection in surveillance videos. ACM Computing Surveys, 2023. **55**(10): p. 1-44.

[8]. Omarov, B., et al., A Skeleton-based Approach for Campus Violence Detection. Computers, Materials & Continua, 2022. **72**(1).

[9]. García-Gómez, J., et al. Violence detection in real environments for smart cities. in Ubiquitous Computing and Ambient Intelligence: 10th International Conference, UCAmI 2016, San Bartolomé de Tirajana, Gran Canaria, Spain, November 29–December 2, 2016, Part II 10. 2016. Springer.

[10]. Hassner, T., Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. in 2012 IEEE computer society conference on computer vision and pattern recognition workshops. 2012. IEEE.

[11]. Gkountakos, K., et al. A crowd analysis framework for detecting violence scenes. in Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020.

[12]. Zhou, P., et al. Violent interaction detection in video based on deep learning. in Journal of physics: conference series. 2017. IOP Publishing.

[13]. Ye, L., et al., Campus violence detection based on artificial intelligent interpretation of surveillance video sequences. Remote Sensing, 2021. **13**(4): p. 628.

[14]. Cheng, M., K. Cai, and M. Li. RWF-2000: an open large scale video database for violence detection. in 2020 25th International Conference on Pattern Recognition (ICPR). 2021. IEEE.

[15]. Zhang, P., et al., View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2019. **41**(8): p. 1963-1978.

[16]. Sudhakar, R. Real Time Violence Detection Using Autonomous Intelligent Surveillance Robot. in 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). 2023. IEEE.

[17]. Kaur, G. and S. Singh, Violence detection in videos using deep learning: A survey. Advances in Information Communication

Technology and Computing: Proceedings of AICTC 2021, 2022: p. 165-173.
[18]. Zhang, C., et al. Pan: Persistent appearance network with an efficient motion cue for fast action recognition. in Proceedings of the 27th ACM International conference on Multimedia. 2019.
[19]. Sultani, W., C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[20]. Lou, Y., et al. Refining the Unseen: Self-supervised Two-stream Feature Extraction for Image Quality Assessment. in 2023 IEEE International Conference on Data Mining (ICDM). 2023. IEEE.
[21]. Kwon, H., et al. Motionsqueeze: Neural motion feature learning for video understanding. in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. 2020. Springer.