Violence Detection in Public Places Based on YOLOv8 Algorithm

Zhixuan Jin

Shandong University, Fushun, China

202200120034@sdu.edu.cn

Abstract. With the acceleration of urbanization and the enhancement of population mobility, the population density has increased, and the proportion of mobile population has risen. The security management of public places faces great challenges. Violence detection based on deep learning algorithms can effectively solve the problems of high labor costs and low monitoring efficiency existing in manual monitoring. In this paper, we propose a method for detecting violence in public places based on the YOLOv8 algorithm, in order to realize the transformation from passive response to active early warning. Experimental results show that the model has a high accuracy rate in violence detection, and the average accuracy rate can reach 0.753 in a comprehensive view, which can effectively reduce labor costs and subjective misjudgment. This paper combines artificial intelligence technology with public security needs,making up for the shortcomings of traditional security through intelligent methods, which is of great significance for building intelligent cities and guaranteeing the security of social governance.

INTRODUCTION

With the acceleration of urbanization and the increase in population mobility, the security management of public places (such as subway stations, shopping malls, etc.) is facing huge challenges. Traditional security systems rely on manual monitoring, which has problems such as high labor costs and low monitoring efficiency. According to the report, the proportion of the global urban population will exceed 68% by 2050. This trend of rapid urbanization makes the security pressure of managing public places increase day by day [1].

The traditional security system mainly relies on manual monitoring and passive response means, which have many problems: Firstly, the high cost of manpower, in developed countries, the cost of manpower has accounted for 65% of the total security budget.Secondly, the high rate of misjudgment of the manual monitoring,.Research has shown that people in the continuous viewing of the video for 30 minutes will be significantly fatigued, and the concentration of attention will be obviously reduced. Studies have shown that the misjudgment rate of human monitoring of critical events has reached 45%, which has far exceeded the standard of social demands [2]. This situation makes the research development and application of intelligent security systems become a research hots pot in the global security field, in which multi-modal fusion and computer vision target detection technology are especially the hot research directions.

In the field of multi-modal fusion and cross-scene behavior recognition, a multi-modal anomaly detection framework fusing visual, audio, and motion features was proposed by the IBM research team for anomalous event recognition in surveillance, which can reduce the false alarm rate by 15% in detection scenarios such as violent acts [3]. In the field of computer vision, the lightweight model based on the YOLO algorithm with edge-end real-time detection is a research hots pot in this field.For example, Korea Goryeo University proposed a lightweight model improved based on YOLOv5, achieving real-time detection at 30FPS and reduced power consumption by 40% [4]. As the latest version of the YOLO series target detection algorithm, YOLOv8 has better model architecture and detection strategy, and the improvement of loss function makes it more adaptable to complex scenarios and has more efficient and wide application value in the field of security monitoring.

This paper describes the principle of the YOLOv8 algorithm, the experimental dataset, the training process, and the experimental results in detail. The purpose of this paper is to train the model using YOLOv8 and test the

practicality of the model in real scenarios. The experiments show that the model has high accuracy and recall rate in violence detection, and can effectively reduce labor costs and subjective misjudgments. Finally, the paper discusses the advantages, challenges, and future research directions of the method.

DATA AND METHODS

Data sources

The dataset used for the experiments is the Violent Flows Data set (http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm) This data set was put forward in 2012 and contains video clips of violent (e.g., fights) and nonviolent scenes (e.g.such as daily activities like walking). The data volume totals 246 video clips (123 violent and 123 non-violent videos), each roughly 15 seconds long. The data is mainly derived from YouTube and public surveillance videos, with street fights, protests, and nonviolent crowd behaviors as the main scenarios. The short time video clips are specialized for violence detection, and are suitable for research in the field of security monitoring.

Methodology

YOLOv8 algorithm

YOLOv8 is the latest version of the YOLO (You Only Look Once) series of target detection algorithms. As the latest iteration of the YOLO series, it has achieved comprehensive improvement in speed and accuracy. Firstly, YOLOv8 adopts an Anchor-free design[5], which directly predicts the center point of the target and bounding box, simplifying the model structure and optimizing the detection ability of small targets. Secondly, YOLOv8 optimizes the CSPDarknet53 backbone[6] network to enhance the feature extraction capability and improve the target differentiation in complex backgrounds.Meanwhile, dynamic label assignment[7] is used to optimize the robustness of overlapping targets in dense scenes and reduce the leakage of detection.Compared with YOLOv5,the YOLOv8 algorithm is also improved in the loss function .YOLOv8's Classification YOLOv8 uses Binary Cross Entropy Loss (BCE Loss)[8] to replace the traditional Softmax, which supports multi-label classification.The regression Loss uses CIoU Loss (Complete IoU)[9] to optimize the location and size prediction of the bounding box and improves the positioning accuracy.YOLOv8 extracts multi-layered features by inputting an image (with the default resolution of 640*640) and extracts the multi-level features[10]using CSPDarknet, then perform shallow and deep feature fusion, directly predicting the bounding boxes, confidence, and category probability. Finally, NMS filters the redundant detection frames and outputs the final detection results.YOLOv8 strikes a balance between speed and accuracy through innovations such as the anchor-free design, dynamic label assignment, and highly efficient backbone network, making it the most advanced real-time target detection algorithm now. Its modular design facilitates developers to optimize for specific tasks.

Experimental Procedures

Using OpenCV, the video case is split into image sequences at a fixed frame rate, and all frames inherit the original video tags fight/no fight. In order to meet the training input requirements of YOLOv8, the image resolution needs to be adjusted to 640*640, and its format needs to be converted to .jpg.Since the dataset lacks the target detection frame labeling, the task needs to be adjusted to image classification (YOLOv8-cls), rather than target detection. For the input single-frame image, the coordinates associated with violence (such as fists, weapons, etc.) are localized to directly decide whether the single frame contains violent behavior. Combining classification loss and regression loss, and using CSPDarknet to extract spatio-temporal features, single-frame reasoning and rapid detection are achieved. Real-time processing of the video stream outputs the probability of violence in each frame. The alarm is triggered when violent behavior is detected in multiple consecutive frames. Meanwhile, in order to prevent overfitting and reasonably evaluate the model performance, the data set needs to be divided into the training set 70%, the validation set 20%, and the test set 10%.

This experiment mainly uses precision rate, recall rate, AP and F1 score to evaluate the model performance. Precision rate is the proportion of results predicted by the model to be in a certain category that really belongs to that category.Recall rate refers to the proportion of the number of targets correctly detected by the model to the total

number of real targets in the data set.AP is the average precision. F1 score is the harmonic average of precision and recall, comprehensively reflecting the classification performance of the model.

ANALYSIS OF RESULTS

As shown in Figure 1, the horizontal axis represents the confidence level of the model on the target detection results, taking the value range of 0-1. The vertical axis represents the accuracy rate. The experimental results show that the accuracy rate increases with the rise of the confidence level, and the accuracy rate for detecting violent behavior is greater than that for nonviolent behavior when the confidence level is low. When the confidence level reaches 0.868, the accuracy rate can reach 1.00. In real scenarios, it is very difficult to have an accuracy of 1.00. This might be due to the sample size of the data set used being too small.

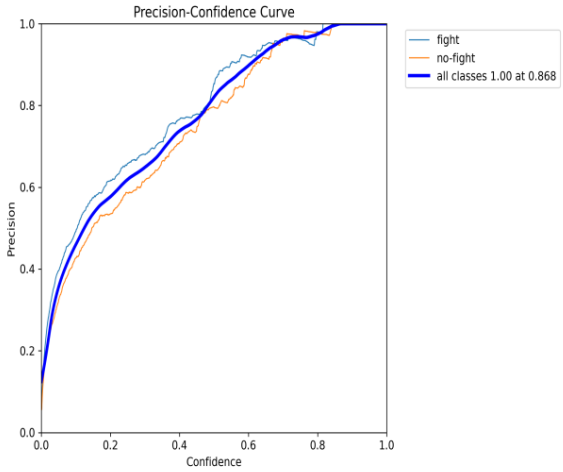

FIGURE 1. Accuracy-Confidence Curve

As shown in Figure 2, the horizontal axis represents the confidence level of the model on the target detection results,with a value range of 0-1. The vertical axis represents the recall rate. The experimental results show that the recall rate decreases with the increase of the confidence level, and the recall rate for violent behavior detection is greater than that for nonviolent behavior when the confidence level is lower. When the confidence level is 0.000, the recall rate can reach 0.92. When the confidence level is set to 0.000, the model makes violent behavior predictions for all inputs, as a consequence, the recall rate reaches 0.92. This indicates that the model's default sensitivity is high, but may lead to an increase in the false alarm rate.

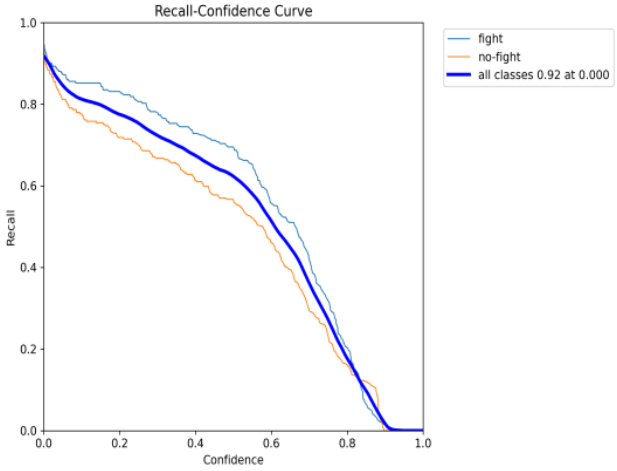

FIGURE 2. Recall-Confidence Curve

As shown in Figure 3,the horizontal axis represents the recall rate and the vertical axis represents the precision rate . The area under the curve represents the AP, which reflects the model's detection performance for this type of targets. It can be seen that the average accuracy of the model for detecting violent behavior is higher than that for nonviolent behavior, and the value of the average precision of the composite curve for all categories is 0.753, which indicates that the model has a high precision for identifying whether it is violent behavior.

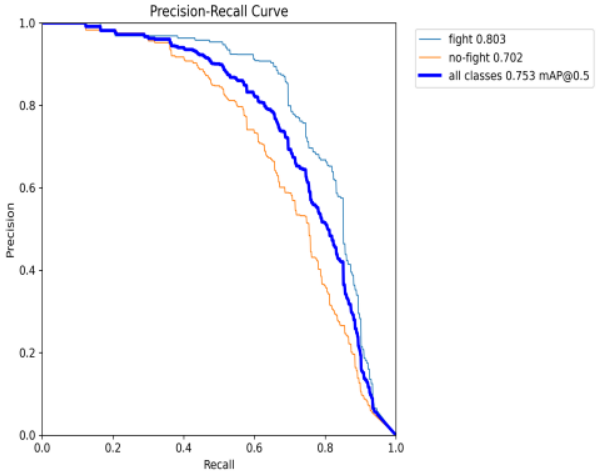

FIGURE 3. Ap Curve

As shown in Figure 4,the horizontal axis represents the confidence threshold of the model (0 to 1) and the vertical axis represents the F1 score . When the confidence threshold is low, the F1 score for violent behavior detection (0.76) is obviously higher than that for nonviolent behavior detection (0.68) at lower confidence thresholds in. The overall F1 score reaches 0.71 at 0.501, which indicates that the precision rate and recall rate of the model were relatively balanced below this threshold.

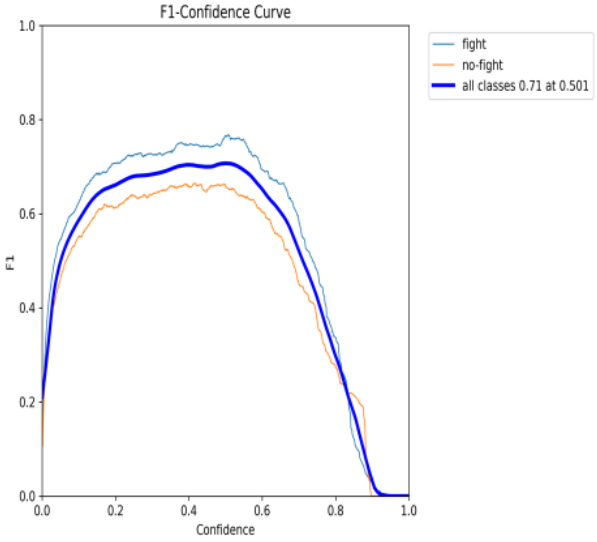

FIGURE 4. Score-Confidence Queering Curve

To sum up, from the above charts, it can be seen that the precision rate will increase with the increase of confidence level, and the recall rate will decrease with the increase of confidence level. However, overall, at a specific confidence level, both the average precision and F1 score have reached a relatively high level, which can be applied to the recognition of violence detection behaviors to greatly reduce the labor cost and subjective misjudgment. At the same time, the confidence threshold of the model can also be changed to make it more consistent with the application scenario. For example, in areas with relatively large flows of people, the confidence threshold can be reduced, in order

to decrease the missed detection rate of the system; in areas with relatively small flows of people, the confidence threshold can be increased to reduce the false alarm rate of the system. However, the experiment also has strong limitations.The current research only relies on the visual information of the video for violent behavior detection, but in real application scenarios, the model performance may significantly decline under the physical conditions of low light or occlusion. The sample size of the data set used in the experiments is relatively small, including only 246 video clips, which cannot cover all real-life violent behavior categories. Most of the violent behaviors in the data set are also simple scenarios, lacking the dynamic interference provided by complex scenes (such as subway stations, concerts, etc.), resulting in a limited ability of the model to generalize to the complex environments. The performance of a single visual modality is limited in scenarios such as low lighting and blurriness. Future research directions can carry out multi-modal data fusion, combined with devices such as infrared cameras, to cover the lack of visual information. For example, thermal imaging can be used to enhance target detection in low-light environments; large-scale multi-modal datasets can also be used to pre-train to enhance the model's ability to understand complex scenarios.

CONCLUSION

This paper uses the Violent Flows Data set to train the model, to study the detection of violence in public places based on the YOLOv8 algorithm and analyzes the model performance by four indexes: precision rate, recall rate, AP, and F1 score. The experiment can show that the YOLOv8 algorithm has application value in the field of security monitoring.By adopting different confidence levels in different scenarios, the model can be more suitable for the application scenarios. The AP of the model can reach 0.753, and the peak of the F1 score can reach 0.71, which is in the middle and upper level of similar studies, and can effectively decrease the labor costs and subjective misjudgments. YOLOv8 has a relatively fast inference speed and is suitable for deployment in devices such as cameras, which is valuable for real-time monitoring in shopping malls, schools, and other scenarios. However, YOLOv8 also has the disadvantage of insufficient recognition ability in low-light, obscured conditions and so on. The model performance can be improved by multi-modal data fusion or diversified data set construction to enhance the wide application and practicality of the YOLOv8 algorithm in the field of security monitoring.

REFERENCES

1. United Nations, Department of Economic and Social Affairs. (2018). *World urbanization prospects: The 2018 revision*. United Nations. https://population.un.org/wup/
2. Schwartz E.; Arbelle A.; Karlinsky L.; Harary S.; Scheidegger F.; Doveh S.; Giryes R. (2024) MAEDAY: MAE for Few- and Zero-Shot AnomalY-Detection DOI:10.1016/j.cviu.2024.103958.
3. Lu Chao, Zeren Zhima, Yang Dehe, Sun Xiaoying, Lv Visiting Xian, Ran Zilin, and Shen Xuhui (2024). Improved YOLOv5 model for lightning whistle acoustic wave lightweight automatic detection https://doi.org/10.11728/cjss2024.03.2023-0067
4. Kim, S., Park, J., & Lee, H. (2022). Lightweight YOLOv5 optimization: achieving real-time detection at 30FPS with 40% power reduction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3), 2876-2884. https://doi.org/10.1609/aaai.v36i3.20222
5. Li Jiachao; Zhou Ya'nan; Zhang He; Pan Dayu; Gu Ying; Luo BinMaize(2024) Plant height automatic reading of measurement scale based on improved YOLOv5 lightweight model DOI:10.7717/peerj-cs.2207
6. Mostafa Farouk Senussi; Hyun Soo Kang(2024)Occlusion Removal in Light-Field Images Using CSPDarknet53 and Bidirectional Feature Pyramid Network: A Multi-Scale Fusion-Based ApproachDOI ： 10.3390/APP14209332
7. Yi Li; Sile Ma; Xiangyuan Jiang; Yizhong Luan; Zecui Jiang(2024)Probability based dynamic soft label assignment for object detection DOI： 10.1016/J.IMAVIS.2024.105240
8. Xinchen Yuan; Xiaojuan Guo; Yande Luo; Xiuhong Guan; Qi Li; Zhiquan Situ; Zijie Zhou; Xin Huang; Zhaowei Rong; Yunhai Lin; Mingxi Liu; Juanni Gong; Hongyan Liu; Qi Yang; Xinchun Li; Rongli Zhang; Chengwang Lei; Shumao Pang; Guoxi Xie(2025)PHNet: A pulmonary hypertension detection network based on cine cardiac magnetic resonance images using a hybrid strategy of adaptive triplet and binary cross-entropy losses.DOI： 10.1109/TMI.2025.3555621
9. Zhang Guoliang; Du Zexu; Lu Weijiang; Meng Xiaoyan(2022)Dense Pedestrian Detection Based on YOLO-V4 Network Reconstruction and CIoU Loss OptimizationDOI： 10.1088/1742-6596/2171/1/012019

10. Hosam S. EL Assiouti; Hadeer El Saadawy; Maryam N. Al Berry; Mohamed F. Tolba(2025)CTRL-F: Pairing convolution with transformer for image classification via multi-level feature cross-attention and representation learning fusionDOI： 10.1016/J.ENGAPPAI.2025.111076