DOI: 10.23977/jaip.2025.080320
ISSN 2371-8412 Vol. 8 Num. 3

Empowering Security Surveillance with Machine Vision: A Survey of Anomaly Detection Technologies

Peng Yin

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China
3473792229@qq.com

Keywords: Machine Vision; Security Surveillance; Anomaly Detection; Object Tracking; UCF-Crime Dataset; Privacy Protection

Abstract: Addressing the pain points of low efficiency, high missed detection rate, and high false alarm rate in manual monitoring for security surveillance, this paper systematically surveys the application of machine vision in anomaly detection. It first sorts out the full-process technical architecture from object detection and tracking to behavior judgment, focusing on analyzing the detection logic of three typical abnormal behaviors: climbing, loitering, and boundary crossing. Secondly, it details the characteristics and application scenarios of two core datasets, UCF-Crime and XD-Violence. Furthermore, it analyzes the impact of complex environments such as insufficient illumination and person occlusion on detection accuracy and corresponding countermeasures. Finally, it discusses the collaborative optimization path between privacy protection and security efficiency, and looks forward to future development directions such as multimodal fusion and common sense reasoning, providing references for the research, development, and implementation of intelligent security systems.

1. Introduction

1.1 Evolution of Technical Requirements for Security Surveillance

With the acceleration of urbanization and the upgrading of public security needs, the traditional monitoring mode relying on manual inspection faces bottlenecks such as limited coverage, delayed real-time response, and high labor costs. Statistics show that when a single monitoring personnel monitors more than 8 video channels simultaneously, the missed detection rate will exceed 40%. Machine vision technology, with its advantages of real-time performance, objectivity, and scalability, has become the core means to break this bottleneck[1]. Through the collaboration of image sensors and algorithm models, it can realize automatic identification and alarm of abnormal behaviors, promoting the transformation of security systems from "post-event tracing" to "pre-event early warning".

1.2 Research Scope and Survey Structure of Anomaly Detection

This paper focuses on three high-frequency and high-risk abnormal behaviors in security scenarios: climbing, loitering, and boundary crossing, and unfolds around the four dimensions of "data-technology-scenario-challenge". Firstly, it elaborates on the general technical process of machine vision detection; secondly, it subdivides the detection links of typical behaviors and verifies them with datasets; then it analyzes the problems of complex environmental interference and privacy protection; finally, it looks forward to technical trends, forming a comprehensive combing of the field of anomaly detection.

2. Core Technical Process of Machine Vision-based Anomaly Detection

2.1 Object Detection: Initial Localization of Abnormal Behaviors

As the foundation of the detection process, this module aims to accurately identify and frame human targets from surveillance video frames. Common algorithms include deep learning-based YOLO series (such as YOLOv5, YOLOv8) and Faster R-CNN. The YOLO series achieves real-time performance of over 30fps through "end-to-end detection", making it suitable for dynamic monitoring scenarios. Faster R-CNN improves localization accuracy through a Region Proposal Network (RPN), and performs better in monitoring static key areas (such as walls and turnstiles). In practical applications, it is necessary to filter target frames through a confidence threshold (usually set to 0.5-0.7) to eliminate low-confidence background interference (such as tree swaying and light changes), providing a reliable target source for subsequent tracking[2].

2.2 Object Tracking: Continuous Capture of Behavioral Trajectories

On the basis of object detection, this module realizes continuous tracking of human targets through inter-frame data association, solving the identity matching problem after targets temporarily leave or are occluded. Mainstream methods include Kalman filtering (predicting target motion trends), Hungarian algorithm (realizing matching between detection frames and trajectories), and SORT (Simple Online and Realtime Tracking) algorithm. Among them, SORT combines motion models with Intersection over Union (IoU) matching, and the tracking accuracy in multi-target crossing scenarios can reach more than 85%. During tracking, a unique ID needs to be assigned to each target, and its timestamp and coordinate information should be recorded to form a complete motion trajectory, providing spatiotemporal data support for behavior analysis.

2.3 Behavioral Feature Extraction: Key Representation of Abnormal Patterns

This module extracts features distinguishing normal/abnormal behaviors from tracked trajectories and target morphologies, which are divided into spatial features and temporal features. Spatial features include human key points (such as 18 joint points extracted by OpenPose: shoulders, elbows, hips, knees, etc.) and target size ratios (such as the ratio of human height to fence height when climbing). Temporal features include movement speed (such as instantaneous speed changes when crossing boundaries) and trajectory curvature (such as statistical analysis of reciprocating trajectory curvature when loitering). In recent years, CNN-LSTM-based hybrid models can automatically extract spatiotemporal fusion features, avoiding the limitations of manual feature design, and the feature representation ability has been improved by more than 20% in detecting complex behaviors (such as multi-person collaborative climbing)[3].

2.4 Behavior Judgment and Alarm: Decision Output of Abnormal Events

This module judges the extracted features based on preset rules or trained models, outputs abnormal results, and triggers alarms. Rule-based methods are suitable for fixed scenarios. For example, boundary crossing behavior is judged by "the number of intersections between the trajectory and the virtual boundary", and loitering behavior is judged by "residence time > T threshold and movement distance < D threshold" (such as T=60s, D=5m). Model-based methods are suitable for complex and variable scenarios. For example, weakly supervised learning-based models (such as MIL multi-instance learning[4]) are trained through video-level labels, which can automatically distinguish abnormal categories such as climbing and fighting. The alarm mechanism needs to support multi-level responses. For example, slight loitering triggers local prompts, and violent climbing triggers sound and light alarms and pushes to managers to ensure the timeliness of emergency disposal[5].

3. Detailed Detection Process of Typical Abnormal Behaviors

3.1 Climbing Detection: From Regional Judgment to Posture Verification

The core features of climbing behavior are "contact between the human body and restricted areas (such as walls and guardrails) + specific postures (such as limb extension and trunk inclination)". The detection process is divided into three steps: first, the system judges whether the human body enters the preset climbing area (such as the 5m range around the wall) through the target tracking trajectory; if yes, the system starts key point detection. Secondly, the system extracts key points of the target such as hips, knees, and ankles, and calculates the angle between the limb connection line and the fence edge (such as a leg-lifting action is judged when the angle between the knee-ankle connection line and the upper edge of the fence is <30°). Finally, the system verifies whether the climbing posture threshold is met for m consecutive frames (such as m=5); if yes, the system triggers a three-level alarm (local pop-up + remote notification). This process needs to solve the occlusion problem. For example, the detection algorithm completes the occluded key points through local feature matching (such as shoulder and neck contours) to ensure the continuity of posture judgment.

3.2 Boundary Crossing Detection: Trajectory Analysis Based on Virtual Boundaries

The core of boundary crossing behavior detection is "the target trajectory crosses the preset virtual boundary". The detection process relies on boundary definition and trajectory intersection judgment: first, the operator sets virtual boundaries (such as straight lines and polygons, supporting one-way/two-way boundary crossing judgment) through the monitoring platform (such as the WEB interface of TP-LINK IPC), and calibrates the mapping relationship between the boundary’s pixel coordinates and actual scales. Secondly, during target tracking, the system real-time calculates the positional relationship between trajectory points and boundaries, and uses the ray casting algorithm to judge whether the target enters the "warning area" from the "safe area". Finally, if the system detects that the trajectory crosses the boundary for 3 consecutive frames and the speed direction conforms to the boundary crossing direction (such as from outside the park to inside), it triggers an alarm. For complex scenarios (such as multi-person parallel boundary crossing), the system needs to distinguish the boundary crossing status of different targets through IDs to avoid misjudging them as a single abnormal event.

3.3 Loitering Detection: Residence Judgment Based on Spatiotemporal Thresholds

The core of loitering behavior is "long-term low-displacement activity of the target in a specific area". The detection process revolves around temporal and spatial thresholds: first, the system obtains the trajectory sequence of each ID through target tracking, and records the initial entry time t_0 and initial position (x_0, y_0). Secondly, the system calculates the current time t_n and current position (x_n, y_n) in real time; if t_n-t_0 > T threshold (such as T=120s, adjustable according to the scenario) and the displacement distance $d = \sqrt{(x_n - x_0)^2 + (y_n - y_0)^2} < D$ threshold (such as D=3m), it marks the target as "suspected loitering". Finally, the system performs trajectory curvature analysis on the suspected loitering target; if the curvature variance > preset value (indicating a reciprocating trajectory), it confirms the loitering behavior and triggers an alarm. In practical applications, users need to exclude normal residence (such as the stay of park security patrols) and filter alarms for specific IDs through a whitelist mechanism to improve detection accuracy.

4. Analysis of Core Datasets for Anomaly Detection

4.1 UCF-Crime Dataset: A Classic Benchmark for Violent Anomalies

As one of the most widely used datasets in the field of anomaly detection, UCF-Crime is constructed by the University of Central Florida. It covers 13 types of abnormal behaviors (such as robbery, fighting, arson, etc.) and normal behaviors (such as walking, standing), including a total of 1900 video clips with a total duration of over 12 hours. The core advantage of this dataset lies in its high annotation quality: each video is annotated with the start/end timestamps of abnormal events, and provides frame-level behavior labels and target bounding boxes. It also supports multimodal analysis; in addition to original RGB videos, it includes frame-level image sequences and optical flow motion data, making it suitable for the training of spatiotemporal feature fusion models. Regarding access, it is not publicly available for free download officially; authorization needs to be applied for through the project homepage or by contacting the author team. Some research institutions provide authorized Baidu Cloud links (note that some violent video samples may be blocked). This dataset is mainly used for algorithm verification of violent abnormal behaviors. For example, the EventVAD model achieves an AUC index of 82.03% on this dataset, which is better than traditional unsupervised methods.

4.2 XD-Violence Dataset: A Large-Scale Breakthrough with Multi-Scenario and Multi-Modal Features

Constructed by the team led by Wu Peng from Xidian University, XD-Violence is one of the largest and most scenario-rich datasets in the current field of anomaly detection. It covers 20 types of violent abnormal behaviors (such as assault, knife attack) and 10 types of normal behaviors, including a total of 8000 videos, covering 15 scenarios such as indoor (such as shopping malls, subways) and outdoor (such as streets, parks). Its core characteristics are reflected in multi-modality and cross-scenario: in addition to RGB data, it includes infrared thermal imaging, depth images and other modalities, which can cope with insufficient illumination scenarios. At the same time, the annotation information includes behavior categories, target IDs, and scenario labels, supporting cross-scenario transfer learning research. This dataset has been open-sourced and widely cited by top conference papers such as ECCV, and is suitable for the training of anomaly detection algorithms in complex scenarios. For example, in high-resolution outdoor scenarios, the detection rate of models trained on this dataset for occluded climbing behavior is 15% higher than that of models trained on UCF-Crime.

5. Technical Challenges in Complex Monitoring Environments

5.1 Impact of Insufficient Illumination and Countermeasures

Insufficient illumination (such as nighttime and rainy weather) will lead to grayscale distortion and reduced contrast of monitoring images, which in turn causes an increase in the missed detection rate of object detection (for example, the grayscale difference between the human body and the background is small at night, and the detection rate of YOLOv5 can drop from 98% during the day to 75%). Current countermeasures mainly include: at the image preprocessing level, using Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast, or separating illumination components and reflection components through the Retinex algorithm to restore target details. To improve the accuracy and robustness of human detection under complex lighting environments, the scheme adopts a systematic design from two aspects: sensor and algorithm. At the sensor level, a multimodal fusion solution of infrared thermal imaging cameras and visible light cameras is employed, utilizing temperature differences in infrared images to achieve precise localization of human targets, thereby compensating for the detection limitations of visible light images in scenarios such as low light and backlighting. At the algorithm level, a feature extraction network with strong illumination robustness (e.g., the improved DarkNet-53) is selected, and illumination disturbance data (such as random brightness adjustment and contrast variation) is incorporated during model training to further enhance the model's adaptability to complex lighting environments. Through the collaborative design of sensor fusion and algorithm optimization, it effectively breaks through the bottleneck of environmental adaptability faced by single-modal detection and traditional algorithms, providing more stable and reliable technical support for human detection tasks.

5.2 Interference of Person Occlusion and Mitigation Measures

Person occlusion (such as crowding and object occlusion) will lead to broken target tracking and missing key points. For example, when the overlap rate of human target detection frames in dense crowds exceeds 60%, the tracking ID switching rate of the SORT algorithm will increase by 40%. Mitigation measures mainly include: at the tracking level, adopting multi-feature fusion matching (such as combining appearance features (HOG) and motion features (velocity vectors)); when the target is occluded, maintaining the tracking trajectory through motion prediction and re-associating after the occlusion is lifted. At the detection level, adopting attention mechanism-based object detection models (such as the attention branch of Faster R-CNN) to focus on unoccluded local areas (such as the head and hands) to realize the recognition of partially occluded targets. At the behavior judgment level, using contextual reasoning (such as judging the overall behavior trend according to the movement direction of unoccluded legs) to make up for the feature loss caused by occlusion and reduce misjudgments caused by incomplete local information.

6. Collaborative Optimization of Privacy Protection and Security Efficiency

6.1 Privacy Leakage Risks and Protection Technologies

When machine vision monitoring collects human behavior data, it is prone to privacy leakage risks (such as the abuse of facial information and activity trajectories). Core protection technologies include: at the data collection stage, the system adopts anonymization processing (such as real-time facial blurring and removal of identity information) and adds noise to data through differential privacy technology to avoid precise positioning of individual information. At the data transmission stage, the system adopts end-to-end encryption (such as TLS 1.3 protocol) to prevent data from being stolen during transmission. At the data storage stage, the system adopts distributed storage and access control (such as the Role-Based Access Control (RBAC) model), allowing only authorized personnel to view sensitive data. At the edge computing level, the system deploys behavior detection tasks on edge devices (such as smart cameras) and only uploads abnormal event results instead of original videos, thereby reducing the transmission and storage of original data and mitigating privacy leakage risks from the source.

6.2 Efficiency Improvement Paths and Trade-off Strategies

The core demand of security efficiency is "real-time detection + rapid response". Improvement paths include: model lightweighting, reducing computational complexity through model pruning, quantization, and knowledge distillation (such as distilling lightweight models from ResNet-50), so that the inference speed of YOLOv8 can be increased by 2-3 times on embedded devices. Hardware acceleration, adopting dedicated computing chips (such as GPU, FPGA, NPU); for example, NVIDIA Jetson Xavier NX can realize real-time detection of 4 channels of 4K videos. System optimization, adopting a hierarchical processing strategy of "coarse detection + fine judgment", first quickly screening suspected abnormal frames through lightweight models, then using high-precision models for fine category judgment to balance accuracy and speed. There is a certain trade-off between privacy protection and efficiency (such as encryption processing increasing computational time). Collaborative optimization of the two can be achieved through hardware-accelerated encryption algorithms and adaptive privacy protection.

7. Conclusion

Machine vision technology has become the core support for anomaly detection in security surveillance. Through the full-process architecture of "object detection - tracking - feature extraction - behavior judgment", it realizes the automatic identification of typical abnormal behaviors such as climbing, loitering, and boundary crossing. Datasets such as UCF-Crime and XD-Violence provide important benchmarks for algorithm research and development. Technical optimizations for insufficient illumination and person occlusion, as well as the collaboration between privacy protection and efficiency, have further promoted the practical application of the technology. In the future, with the development of technologies such as multimodal fusion, common sense reasoning, and adaptive transfer, the application of machine vision in security surveillance will be more intelligent[6], scenario-based, and humanized, providing more reliable technical guarantees for public security, park management, traffic control and other fields.

References

[1] Mohindru, Vandana, and Shafali Singla. "A review of anomaly detection techniques using computer vision." The International Conference on Recent Innovations in Computing. Singapore: Springer Singapore, 2020.
[2] Madan, Neelu, et al. "Self-supervised masked convolutional transformer block for anomaly detection." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.1 (2023): 525-542.
[3] Hasan, Mahmudul, et al. "Learning temporal regularity in video sequences." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[4] Pillai, Gargi V., Ashish Verma, and Debashis Sen. "Transformer based self-context aware prediction for few-shot anomaly detection in videos." 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022.
[5] Yu, Jiawei, et al. "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows." arxiv preprint arxiv:2111.07677 (2021).
[6] Latapie, Hugo. "Common sense is all you need." arxiv preprint arxiv:2501.06642 (2025).