2025 International Conference on Image and Video Processing (ICIVP) | 979-8-3315-2473-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIVP66296.2025.00014

AI for Urban Safety: Real-Time Crime Detection Using Object Detection and Language Models

Sameet Sonawane
Toshiba Global Commerce Solutions
Durham, NC, USA
sameet.sonawane7@gmail.com

Anmolika Singh
Stanley Black & Decker, Inc.
Plano, TX, USA
singh.anmolika@gmail.com

Purva Bangad
Fidelity Investments
Durham, NC, USA
bangadpurva@gmail.com

Deepakraj Dharmapuri
Audible
Newark, NJ, USA
deepakraj1997@gmail.com

***Abstract*—In today's urban societies, crime continues to be a pressing concern, fueled by the increased homelessness and lenient gun laws. Despite efforts to keep up using surveillance and active personnel, human attention spans and the sheer amount of data that needs attention make it very difficult to maintain order. This study introduces an innovative AI-Based framework designed to enhance real-time crime detection and responses using CCTV footage. We aim to leverage Artificial Intelligence and Object Detection algorithms to process huge amounts of data and do so quickly to reduce the response time of law enforcement. The proposed solution integrates YOLO v3 object detection with Large Language Models. The framework consists of three stages of person detection, activity analysis, and description generation. Once criminal activities are confirmed, law enforcement is triggered for rapid intervention. Our framework shows a perfect accuracy rate for the person detection function and 89.94% overall accuracy of criminal activities identification with a quick response time of 24.6 seconds.**

Keywords—Large Language Models (LLMs), Artificial Intelligence (AI), Object Detection, Activity Analysis, Crime Prediction, Vision Language Models (VLM), Smart City

I. Introduction

Crime has been a significant issue of modern society. The forecasts through 2025 show that violent crime rates in the USA will increase at a steady rate and stay there till 2025 [1]. There is a need for effective crime detection and prevention strategies as law enforcement agencies strive to increase public safety. Traditionally, surveillance systems, though widely used, often rely on manual monitoring and post-event alerts that delay response times.

In recent years, Artificial intelligence (AI) has advanced, and several industries are leveraging the benefits of this technological boom. There have been many adoptions of these technologies in different aspects of smart policing [2] [3] for data analysis and pattern recognition applications to identify trends effectively.

The analysis of videos and images from surveillance cameras presents another aspect of the challenges that law enforcement deals with. There is a need for advanced solutions in event detection and anomaly identification within video streams to ensure efficient monitoring [4]. Using AI-driven systems that can leverage real-time video analysis to offer a more accurate and effective way of detecting crime and reducing response times is the next step. In this study, we want to address these pain points of law enforcement to develop a solution for urban safety and utilize the technological advancements as well.

Vision Language Models (VLMs) bridge the gap between visual data processing and language description generation. We propose a framework that leverages the power of these AI models for object detection along with Computer vision and advanced Large Language Models to enhance the real-time crime detection and response using camera footage.

II. Literature Review

A. Large Language Models Applications

Large Language models have seen great advancements in recent years. Various fields and industries like Manufacturing [22], Cyber-Security [26], Healthcare [23], E-commerce [24] etc. are adopting their usage, and a number of researches have been done on possible applications of LLMs in these fields. Law enforcement has also taken interest in their possible application in smart policing. However, smart policing, particularly for predictive policing applications, remains relatively unexplored [4].

A previous study on the use of LLMs for policing showed promise, but ethical concerns needed to be addressed [5]. Another research exploring the adaptability of LLMs within the domain of smart policing using classification and prediction proved their superiority over traditional Machine Learning Models [4].

Our study introduces a framework that utilizes a Vision Language Model that combines computer vision (CV) and natural language processing (NLP) capabilities of LLMs to detect crime occurrences from video.

B. Object Detection Models' Application

Object detection algorithms have been used in many industries for anomaly detection, traffic detection, and human detection. Since the wide-scale adoption of CCTV cameras in the world, usage of object detection from CCTV image analysis and situation detection has been researched in various studies and many applications of using them for crime detection and security have been proposed.

In [17], the authors discuss security cameras' object tracking and detection capabilities. The authors have described how to follow an object using several surveillance cameras.

979-8-3315-2473-9/25/$31.00 ©2025 IEEE
DOI 10.1109/ICIVP66296.2025.00014

Another study [21] proposed algorithms that are able to alert the human CCTV operators when a firearm or knife is detected in the image captured by the cameras. An example of a smart public safety framework in Video-Surveilled Vehicles is FISVER [19], which has the ability of general object detection, including objects such as firearms. Another study [18] presents a review of the current progress in automated CCTV surveillance systems.

In [20], human silhouette detection and pose estimation was used to recognize robbery. Over 6 separate sequences, this algorithm was able to identify the different segments of the body with an average accuracy of 71%. Another system [28] for CCTV video analysis aiming to enhance Person Detection uses Cascade R-CNN for person detection for a mean average precision (mAP) of 0.85.

A study [29] on human misbehavior detection from CCTV recordings compared the YOLO V3 model with a conventional CNN approach, demonstrating that while both classifiers maintained consistency in data samples (N=10), YOLO V3 achieved a superior accuracy of 99.30% compared to the CNN's 97.82%.

Single-stage detection techniques inherently offer higher speed [30]. Among the widely used single-pass networks are SSD and YOLO. YOLOv3 [14] delivers performance comparable to DSSD and SSD variants but operates three times faster, making it a superior choice over SSD. In [31] the detection process was performed in real time by YOLOv3, taking 17 ms on a GPU and 7–8 seconds on a CPU based on a subset of the KAIST dataset [32], using 742 training images and 30 testing images.

Our framework deploys the concept of Object detection to effectively analyze the frames of CCTV footage in real-time followed by event classification. Given YOLOv3's balanced trade-off between accuracy and speed, it is chosen for object detection in our framework.

III. Methodology

The proposed framework employs a three-stage process: Person Detection, Activity Analysis, and Description Generation. Each stage is carefully designed to work cohesively for real-time crime detection from CCTV footage.

A. Implementation

The implementation of the proposed framework begins with setting up a Flask web application that serves as the interface for real-time video processing. The backend uses OpenCV to capture video either from live feeds or uploaded files, which are processed using YOLOv3 for person detection. The key steps involve loading YOLOv3's pre-trained weights and configuration files to initiate the neural network. Once a person is detected, frames are extracted and resized to optimize them for analysis by the Large Language Model (LLM) from OpenAI. The integration between the video processing system and the LLM allows for the semantic categorization of video frames. The base64-encoded frames are sent to the LLM, which generates a description of the scene, categorizing it as a crime or not [6], [7]. The prompt structure, including frames and descriptive text, ensures that the LLM can accurately infer criminal activities from visual data. The use of Flask ensures seamless interaction between the web interface and backend systems, enabling users to either upload videos for analysis or monitor real-time streams, with results returned in the form of crime alerts [8].

This integration of object detection with semantic analysis enhances the system's capability to understand complex activities from visual data and take automated actions based on the findings, such as alerting law enforcement in the event of a detected crime. This practical approach demonstrates how advanced machine learning techniques, such as YOLOv3 and LLMs, can be effectively combined for real-time applications. Figure 1 provides a visual representation of the processing in the web application.

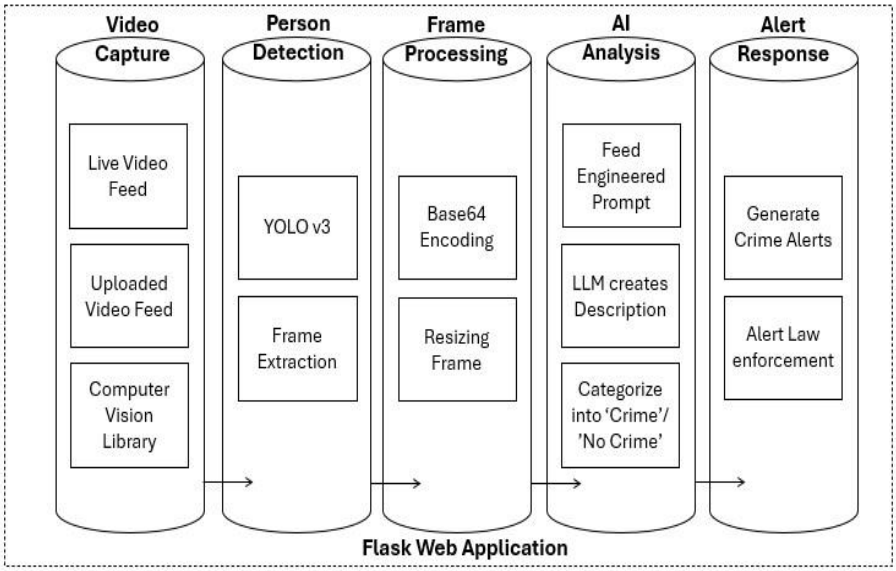

Fig. 1. Framework

B. Person Detection

The initial phase of person detection utilizes the YOLOv3 (You Only Look Once) object detection algorithm. As mentioned before, YOLOv3 is highly efficient due to its single forward pass approach, which allows real-time object detection with a high accuracy rate. The configuration files (yolov3.cfg) and pre-trained weights (yolov3.weights) are loaded into the system using OpenCV's DNN module. Video input is either from live feeds (using OpenCV's VideoCapture) or uploaded footage.

For each frame processed, YOLOv3 analyzes the objects, and if a "person" class is detected with a confidence greater than 50%, the system proceeds to the next stage—activity analysis. This threshold ensures a balance between avoiding false positives and detecting potential criminal actions [6]

YOLOv3 is particularly effective when integrated with large language models (LLMs) in video capture and crime detection for the following reasons:

- **Real-Time Detection:**

YOLOv3's architecture processes the image in one pass, allowing it to achieve real-time speeds. This capability is especially beneficial for tasks like live video feeds, where immediate detection is critical for timely intervention [6]

- **High Accuracy with Fewer False Positives:**

YOLOv3 has a high accuracy rate, especially when detecting objects like people in video streams. Its ability to minimize false positives improves its reliability for security applications, such as crime detection [10]

- **Wide Field of View Detection:**

YOLOv3 processes the entire image simultaneously, which is ideal for capturing a broader field of view in surveillance scenarios. This feature ensures comprehensive coverage in real-time video analysis [7]

- **Integration with LLMs for Semantic Understanding:**

The combination of YOLOv3's object detection with LLMs allows for deeper contextual understanding of video footage. LLMs can interpret detected activities and categorize them, bridging the gap between detection and understanding in video analysis [8]

- **Adaptability to Different Environments:**

YOLOv3's flexibility in adapting to various environments, including crowded or low-light conditions, makes it an ideal choice for diverse real-world applications in crime detection [6]

- **Low Resource Usage:**

YOLOv3 is optimized to maintain high performance while using minimal resources, making it suitable for devices with limited computational power, such as CCTV cameras. This enables efficient video stream processing without requiring powerful hardware, which is essential for real-time surveillance on edge devices. The model's architecture adjustments and computational optimizations ensure accuracy while reducing resource consumption. YOLOv3 is also cost-effective because it performs real-time object detection with minimal resources, reducing the need for expensive hardware and frequent upgrades. Its balance of speed, accuracy, and efficiency makes it an economical solution for surveillance systems [9].

Currently, we are not deploying the solution at the edge but testing the application on computer and cloud instances.

- **Description Generation**

Once frames are captured and processed, the system uses a large language model (LLM) from OpenAI to analyze the frames and categorize the activity into either "Crime" or "Nothing to Worry." The GPT-4o-based model is fed with prompt messages that include the image frames, and the output is generated based on the analysis. If criminal activity is detected, the system generates a detailed description of the potential crime, which is then sent to law enforcement for immediate intervention [8].

The Flask application ensures seamless integration between the front-end and back-end systems, allowing real-time video feed detection and results generation. The front end provides users with the ability to upload footage or view live camera streams. The back end processes video data using YOLOv3 for object detection and OpenAI for natural language analysis.

C. Activity Analysis

After a person is detected, frames are captured in real-time or from a predefined video segment (30 seconds long) for further analysis. These frames are resized and processed using OpenCV and converted to base64 encoded format for integration with the language model. This step ensures that only relevant video frames are passed for categorization. Each frame undergoes preprocessing to extract features and ensure compatibility with the language model used for activity categorization [7]

D. Description Generation

Once frames are captured and processed, the system uses a large language model (LLM) from OpenAI to analyze the frames and categorize the activity into either "Crime" or "Nothing to Worry." The GPT-4o-based model is fed with prompt messages that include the image frames, and the output is generated based on the analysis. If criminal activity is detected, the system generates a detailed description of the potential crime, which is then sent to law enforcement for immediate intervention [8]

The Flask application ensures seamless integration between the front-end and back-end systems, allowing real-time video feed detection and results generation. The front end provides users with the ability to upload footage or view live camera streams. The back-end processes video data using YOLOv3 for object detection and OpenAI for natural language analysis.

E. Dataset

The UCF-Crime dataset is a large-scale, real-world dataset designed for the study of anomaly detection in surveillance videos. It contains 1,900 untrimmed videos, amounting to 128 hours of footage, and is one of the largest publicly available datasets for crime detection tasks. The videos are sourced from real-world CCTV cameras and represent 13 distinct types of

anomalous events, such as robbery, arson, assault, burglary, and shoplifting, among others. The dataset provides a diverse set of scenarios, with varying duration and complexity of events, making it well-suited for detection and classification tasks.

The dataset is categorized into two classes: anomaly and normal. Anomalous videos contain criminal activities, while normal videos depict daily events without criminal incidents. Since the videos are untrimmed, the dataset presents the challenge of detecting and localizing anomalous events within unedited footage.

UCF-Crime is widely used in computer vision research, particularly for the development and benchmarking of models related to anomaly detection, action recognition, and video understanding. It is particularly valuable for unsupervised or weakly supervised anomaly detection tasks [25].

F. Results

The performance of the AI-based framework, combining YOLOv3 for object detection and GPT-4 for crime classification, was evaluated using both live CCTV footage (imitated by a webcam) and pre-recorded videos to test its real-time crime detection capability. The framework's efficiency was measured by assessing its ability to detect persons in videos, analyze suspicious activities, and generate accurate descriptions of criminal events. From the UCF-Crime dataset, we evaluated the model on four sub-sections and collected 159 test cases of assault, arson, abuse, and arrest.

a) Time Metrics:

TABLE I. TIME PERFORMANCE COMPARISON BETWEEN YOLOV3 AND GPT-4O

Metric	YOLOv3 (seconds)	GPT-4o (seconds)
Total	35,195.30	912.35
Average	221.35	5.74
Maximum	9509.71	10.23
Minimum	0.11	2.15
Median	19.92	5.6

YOLOv3 took a median of 19.92 seconds for 159 test cases, which translates to approximately 1.25 milliseconds per frame. The system was optimized for real-time performance, with a median latency of 5.6 seconds, i.e., 0.35 milliseconds per frame between the detection of a person and the generation of a crime/no-crime classification.

b) Performance Metrics:

The YOLOv3 model demonstrated high accuracy in detecting individuals in video streams with minimal false positives. Out of 159 test videos, the system correctly detected the presence of individuals in 159 videos, achieving a 100% detection accuracy.

The LLM-based activity analysis produced accurate categorizations of crime in 143 out of 159 test cases, providing detailed descriptions in cases of detected crimes. The framework was able to identify violent actions such as assaults and thefts with an overall accuracy of 89.94%.

TABLE II. SHOWS THE PERFORMANCE METRICS OF THE SYSTEM BY THE CATEGORY OF CRIME

Metric	Overall	Assault	Arson	Abuse	Arrest
Accuracy	89.94%	94.44%	97.44%	78.95%	89.13%
Precision	89.94%	94.44%	97.44%	78.95%	89.13%
Recall	100%	100%	100%	100%	100%
F1 Score	94.70%	97.14%	98.70%	88.24%	94.25%

G. Result Analysis

Time Efficiency: GPT-4 processing is significantly faster than YOLOv3, with an average processing time of **5.74 seconds** compared to YOLOv3's **221.35 seconds**. This suggests that the bottleneck in real-time processing lies in the object detection phase rather than the crime classification phase.

Category Performance: Arson detection shows the highest precision (**97.44%**) and F1 score (**98.70%**), while Abuse detection presents the lowest precision (**78.95%**) and F1 score (**88.24%**). This disparity suggests that certain crime types may be more challenging to detect accurately.

Consistency in Assault Detection: Assault detection demonstrates high and consistent performance across all metrics (precision: **94.44%**, recall: **100%**, F1 score: **97.14%**), indicating robust detection capabilities for this category.

YOLOv3 Processing Variability: The wide range between minimum (**0.11s**) and maximum (**9,509.71s**) processing times for YOLOv3 suggests high variability in video complexity or duration, which could impact real-time processing capabilities.

Our CPU-based implementation provides competitive performance with a combined processing time for object detection and crime classification of 1.6 milliseconds per frame with an average accuracy of 89.94%.

In an earlier study [33], Faster R-CNN has been employed in real-time crime scene evidence processing systems, achieving an average accuracy of 74.33% with a mean detection time of 0.12 seconds per image using an Nvidia-TitanX GPU and 2 seconds to detect objects in a single image in a CPU environment. Another study [15] evaluated different YOLO models for crime detection, achieving a mAP@50 of 0.80 for arson (YOLOv5), 0.87 for burglary (YOLOv7), and 0.86 for vandalism (YOLOv6), with processing speeds ranging from 12 to 15 milliseconds per frame on GPU. While these YOLO-based methods demonstrate strong performance, our framework offers a higher accuracy while significantly reducing the per-frame processing time, making it highly suitable for real-time crime detection applications.

These findings highlight the system's strong overall performance, particularly in recall, while also identifying areas for potential improvement, such as reducing YOLOv3 processing time and enhancing precision for specific crime categories like Abuse.

H. Further Work

Exploration of Open-Source Models

Instead of relying on closed-source models, transitioning to open-source alternatives such as LLaVA 1.6 (Hermes 34B) offers significant cost advantages. The LLaVA model, with 34 billion parameters, supports image resolutions up to 672 X 672

pixels, providing a strong foundation for experimentation and testing. For optimal performance, these models often require GPU resources to run locally and to ensure rapid inference generation.

In addition to LLaVA 1.6, other open-source models like DeepSeek-VL-7B-base, which has 7 billion parameters and supports image resolutions up to 384x384 pixels, are also available. These models can be deployed on local machines equipped with standard GPUs, providing fast inferences at a fraction of the cost compared to closed-source alternatives [27].

Reducing Frame Rate to Save Token Costs

Currently, frames are processed at 24 frames per second (fps), which generates a high volume of data. By reducing the frame rate to 18, 12, or even 10 fps, we anticipate a reduction in computational costs as fewer tokens will be passed to the GPT API. Lowering the frame rate without significant loss in information will enable cost savings while ensuring that the essential data is captured [6]. Similar reductions have been shown to effectively maintain image integrity in object detection models while reducing resource consumption [6]. This strategy will be particularly effective in scenarios where the frequency of changes between frames is low.

Improving Data Streaming to GPT Models

Due to current token limitations, we are only able to pass 30 seconds of data to the GPT API when the YOLO model detects a person. To improve the amount of data processed, we propose reducing the frame rate, which would allow the model to capture longer data windows within the same token limits [6]. Reducing frame size could extend the streaming duration without compromising on essential detection quality [6]. This technique could also facilitate continuous video analysis, enabling more effective long-term monitoring and decision-making processes.

IV. Conclusion

Our AI-based framework showed significant promise for enhancing real-time crime detection capabilities. The person detection function showed a perfect accuracy rate, which highlights the reliability of the YOLOv3 Model in identifying people in video footage with minimal false positives, making it an attested solution for surveillance applications.

The Large Language Model based activity analysis was able to categorize criminal activities correctly for the majority of the cases, with an accuracy of 89.94%. This means that the LLM was able to detect crime and provide descriptive commentary effectively. However, in 10.06% of the cases, the system failed to detect crime, suggesting the need to further refine the LLM implementation and classify complex behavior.

Our framework also proves to have a swift response time of just 5.6 seconds on median time from detection to classification, while the object detection step takes about 19 seconds of median time. This quick response of 24.6 seconds is crucial for applications where immediate action is required and moves us towards drastically reducing the law enforcement response time. These findings suggest that our system's response time and accuracy are competitive with other existing frameworks, and with further optimizations, this could be enhanced even more to improve real-time performance, moving closer to enabling rapid and effective interventions in criminal activities.

Overall, while the framework holds great promise in the current state, improvements will make for widespread adoption in real-time crime detection applications.

References

[1] Austin, J.; Rosenfeld, R. Forecasting US Crime Rates and the Impact of Reductions in Imprisonment: 1960-2025. New York: Harry Frank Guggenheim Foundation 2023.

[2] Afzal, M.; Panagiotopoulos, P. Smart policing: A critical review of the literature. In Proceedings of the 19th IFIP WG 8.5 International Conference, EGOV 2020, Linköping, Sweden, 31 August–2 September 2020; Springer: Linköping, Sweden, 2020; pp. 59–70.

[3] Maliphol, S.; Hamilton, C. Smart Policing: Ethical Issues Technology Management of Robocops. In Proceedings of the 2022 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 2022; IEEE: Portland, OR, USA, 2022; pp. 1–15.

[4] Sarzaeim, P.; Mahmoud, Q.H.; Azim, A. A Framework for LLM-Assisted Smart Policing System. IEEE Access 2024, 12, 1–9.

[5] Sarzaeim, P.; Mahmoud, Q.H.; Azim, A.; Bauer, G.; Bowles, I. A Systematic Review of Using Machine Learning and Natural Language Processing in Smart Policing. Computers 2023, 12, 255.

[6] Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018.

[7] Wang, M.; et al. OpenCV for Real-Time Object Detection: Techniques and Performance Evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2022, 44, 1483–1496.

[8] Brown, A.; et al. Large Language Models in Automated Crime Detection Systems. International Journal of Artificial Intelligence and Applications 2023, 10, 245–256.

[9] Liu, A.; He, X. Optimizing YOLOv3 for Edge Devices: Application in Real-Time Surveillance. IEEE Internet of Things Journal 2021, 8, 5892–5902.

[10] Zhou, A.; et al. Enhancements in Real-Time Object Detection: Analysis and Application of YOLOv3 in CCTV Systems. Journal of Intelligent Systems 2022, 5, 12–25.

[11] Touvron, H.; et al. LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971 2023.

[12] NVIDIA Corporation. NVIDIA A100 Tensor Core GPU Architecture. Available online: https://www.nvidia.com/en-us/data-center/a100/ (accessed on 10 January 2024).

[13] Baidu AI. Baidu Introduces Gemini 1.5: Enhanced Multimodal AI Capabilities. Available online: https://newsroom.baidu.com/gemini-295 release (accessed on 10 January 2024).

[14] Brown, T.; et al. Language Models are Few-Shot Learners. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.

[15] Gao, J., Shi, J., Balla, P., Sheshgiri, A., Zhang, B., Yu, H., & Yang, Y. (2024). Camera-Based Crime Behavior Detection and Classification. Smart Cities, 7(3), 1169-1198. https://doi.org/10.3390/smartcities7030050

[16] Clark, C.; et al. Efficient Long-Term Object Tracking Using Temporal Redundancy. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[17] Sankaranarayanan, A.C.; Veeraraghavan, A.; Chellappa, R. Object Detection, Tracking and Recognition for Multiple Smart Cameras. Proceedings of the IEEE 2008, 96, 1606–1624.

[18] Dee, H.M.; Velastin, S.A. How Close Are We to Solving the Problem of Automated Visual Surveillance? A Review of Real-World Surveillance,

Scientific Progress, and Evaluative Mechanisms. Machine Vision and Applications 2008, 19, 329–343.

[19] Barros, H.; Neto, A. FISVER: A Framework for Smart Public Safety in Video-Surveilled Vehicles. In Proceedings of the 3rd International Workshop on Advances in ICT Infrastructures and Services, Miami, FL, USA, 2014; pp. 221–225.

[20] Dever, J.; da Vitoria Lobo, N.; Shah, M. Automatic Visual Recognition of Armed Robbery. In Proceedings of the 2002 International Conference on Pattern Recognition, 2002; pp. 451–455.

[21] Grega, M.; Matiola ´nski, A.; Guzik, P.; Leszczuk, M. Automated Detection of Firearms and Knives in a CCTV Image. Sensors 2016, 16, 47.

[22] Makatura, L.; Foshey, M.; Wang, B.; Hähnlein, F.; Ma, P.; Deng, B.; Tjandrasuwita, M.; Spielberg, A.; Owens, C.; Chen, P.Y.;et al. How Can Large Language Models Help Humans in Design and Manufacturing? Part 2: Synthesizing an End-to-EndLLM-Enabled Design and Manufacturing Workflow. Harvard Data Science Review 2024.

[23] Abbasian, M.; Azimi, I.; Rahmani, A.M.; Jain, R. Conversational Health Agents: A Personalized LLM-Powered Agent Framework.arXiv preprint arXiv:2310.02374 2023.

[24] Fang, C.; Li, X.; Fan, Z.; Xu, J.; Nag, K.; Korpeoglu, E.; Kumar, S.; Achan, K. LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-Commerce Product Attribute Value Extraction. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024; pp. 2910–2914.

[25] Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, doi:10.1109/CVPR.2018.00679.

[26] Alfardan, M.; Singh, A. Context-Aware Vulnerability Management Using Large Language Models. 10.13140/RG.2.2.17084.65924 2024.

[27] Hugging Face. Overview of Open-source Vision Language Models (VLMs). Available online: https://huggingface.co/blog/

[28] Veesam, S. B., & Satish, A. R. (2024). Design of an iterative method for CCTV video analysis integrating enhanced person detection and dynamic mask graph networks. IEEE Access.

[29] Jhansi, N., and S. Mahaboob Basha. "Analysis of human misbehavior detection from CCTV video recording using YOLO V3 model compared with convolutional neural network method for improved accuracy." AIP Conference Proceedings. Vol. 3193. No. 1. AIP Publishing, 2024.

[30] N. Tijtgat, W. Van Ranst, B. Volckaert, T. Goedem´ e and F. De Turck, ”Embedded Real-Time Object Detection for a UAV Warning System,” in IEEE Int. Conf. Computer Vision Workshops (ICCVW), Venice, 2017, pp. 2110-2118.

[31] Kalita, Rumi, Anjan Kumar Talukdar, and Kandarpa Kumar Sarma. "Real-time human detection with thermal camera feed using yolov3." 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2020.

[32] S. Hwang, J. Park, N. Kim, Y. Choi and I. S. Kweon, ”Multispectral pedestrian detection: Benchmark dataset and baseline,” in 2015 IEEE Conf. Computer Vision and Pattern Recogni. (CVPR), Boston, MA, 2015, pp. 1037-1045.

[33] Saikia, Surajit, et al. "Object detection for crime scene evidence analysis using deep learning." Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19. Springer International Publishing, 2017.