AutoWatcher: a Real-Time Context-Aware Security Alert System using LLMs

Praneeth Vadlapati
University of Arizona,
Tucson, USA
praneethv@arizona.edu
ORCID: 0009-0006-2592-2564

Abstract: Ensuring the safety of every person is a necessity. Burglaries are frequent across residential and commercial spaces. Criminals frequently use weapons and are a serious threat to the common people who are innocent. The best solution to safety is alertness. Computer Vision (CV) is frequently implemented through neural networks that use a process of training to gain abilities to detect humans. Multimodal Large Language Models (LLMs) capable of vision-based tasks are common and are capable of processing images similar to humans who use their analytical abilities. With the ongoing research, LLMs are becoming smaller, faster, more accurate, and cost-effective, making them usable in security systems to detect threats in an intelligent way based on custom instructions. Existing research does not explore the usage of multimodal LLMs for real-time monitoring. This paper proposes a system called "AutoWatcher" to monitor a place using a camera in real-time, constantly detect humans, and use LLMs to assess potential threats. The system is designed for two levels of alerts, one when a person is detected and another when the LLM declares them as suspicious. This ensures alertness of the people who can proactively protect their lives and valuable assets on time before a security breach occurs at their home or workspace. This enables the residents to alert the emergency service authorities in their locality. The system has successfully detected suspicious people with an accuracy of up to 90%. The accuracy is 100% in some test cases. The code is available at github.com/Pro-GenAI/AutoWatcher.

Keywords: large language models (LLMs), computer vision (CV), multimodal LLMs, neural networks, edge AI, automated surveillance

I. INTRODUCTION

Maintaining safety and security in public and private places is paramount, considering the concerns regarding growing burglary and crime rates in residential and commercial spaces across the world [1], [2]. Traditional surveillance systems are useful in recording several places [3]. However, the systems lack the identification of threats in real-time to alert the potential victims [3].

A. *Disadvantage of existing approaches*

The existing work on motion detection [4] lacks the capability to differentiate between potential threats and the normal presence of humans. Hence, the alerts generated using the existing work necessitate constant observation by security personnel, which is not affordable for non-profit organizations and people in residential areas. To mitigate threats and surveillance issues, there is a growing requirement for intelligent surveillance systems capable of immediate threat detection and alerting [5].

Computer Vision (CV) technology based on Neural Networks (NNs) experienced major breakthroughs in this domain, using intelligent NNs to detect humans [6], [7], [8]. However, these systems require training data, time, and resources for the training process. NNs lack an understanding of differentiating false alarms and real threats. For example, the existing systems might detect humans who walk on the road and assume them as suspicious. A potential solution is offered by recent advancements in multimodal large language models (LLMs) that can process images based on instructions provided as text [9], in a manner similar to humans making inferences based on context and patterns.

B. Proposed system and its benefits

This paper proposes a system called “AutoWatcher,” a non-conventional system that integrates CV with multimodal LLMs to monitor using surveillance cameras in real time. The system leverages an LLM to analyze the captured footage when detecting human presence, to know whether the person who appears to the camera is a real threat, and to minimize false alarms that disturb the users who otherwise deserve to relax. This ensures a reduction in user inconvenience. The system is designed for two levels of alerts, including an initial alert when a person is detected and a main alert when a person is classified as suspicious. This approach aims to provide an initial alert on time, which is useful to ensure safety in cases where the users get ready to open the door to go out. The decisions generated by the LLM include a reason, enabling explainability and transparency of the decisions generated by the system, which helps the users understand the reason for each alert and facilitates constant improvements of the system.

II. RELATED WORK

Significant research exists on the integration of artificial intelligence (AI) in security systems. CV-based systems have been used in surveillance to detect objects, faces, and even criminal activities. For instance, object detection models like YOLO [6], [7] and faster R-CNN [10] were widely adopted to identify humans in real-time in continuous video streams. However, these models are primarily focused on object identification and lack the capability to have a contextual understanding of the streams. LLMs use a contextual understanding and follow the instructions mentioned using custom instructions [9]. The CLIP model [11] by OpenAI demonstrated the ability to understand and relate text descriptions to images. Flamingo by Google Deepmind [12] has explored multimodal learning to process visual data with language. Despite these advances, there is a research gap in the application of multimodal LLMs for real-time security surveillance using custom instructions.

Febriantono et al. (2023) [13] proposed a smart home security system using face recognition based on CNN. Mukto et al. (2024) [14] worked on a real-time crime monitoring system using deep learning techniques. The work included behavioral pattern analysis for crime prediction. Olszak (2024) [15] worked on enhancing security with LLMs and CV to constantly use an LLM to describe an image to enable decisions for further actions. However, it does not focus on detecting motion and humans in real-time and using an LLM using custom instructions only when required. The existing work did not include context-aware decision-making based on LLMs. Existing work focuses on human detection and image analysis using AI, with a research gap in integrating both real-time monitoring and threat evaluation through LLMs using custom instructions. This paper uses these advancements to propose a hybrid system that uses CV to detect movements, YOLO to detect humans, and LLMs to detect threats in real-time.

III. METHODS

A. Selecting models

For the first level alert of human detection, the object detection model YOLO (v8) [6], [7] is used, considering its ability to detect humans in various scenarios, including blurred images and low light. Multiple multimodal LLMs capable of vision tasks were used for effective experimentation. Selecting multiple LLMs allows to experiment on the effectiveness of the system as well as the effectiveness of each LLM in the system. The multimodal models selected are GPT 4 Turbo [17], which is proprietary, and LLaVA-Phi 3 [18], which is an open-weight model. These models are selected based on their ability to understand images.

B. Selecting sample videos

A diverse set of videos was carefully curated from different sources to simulate a wide range of scenarios. The videos vary in lighting conditions to test the robustness of the system under various scenarios, including low light.

Multiple types of videos selected from multiple sources are downloaded into the system for further processing. The videos included are about suspicious people peeking into a house [19], peeking into a window [20], two thieves stealing from a house [21], a person pointing a gun at a door [22], and a thief stealing a bag from a car [23].

C. Optimizing the file size of videos

The videos are pre-processed using FFmpeg [24] to reduce the resolution to 640p*360p. The frame rate has been reduced to 30 frames per second. This ensures that on-device processing is allowed on devices with fewer resources. Such reduction ensures a reduction in file sizes without losing important details. Videos with fewer file sizes are faster to process when the resources are low. Handling multiple video streams simultaneously is possible due to a reduction in file sizes. Smaller images consume less time for the LLMs to process.

D. Motion detection

OpenCV [4], [16] is utilized for motion detection in the video feed in real-time, serving as an initial step to identify potential frames with movements. The process uses a background subtraction technique to distinguish the movements in the footage from a static background. If a motion is detected, such as a human entering, the frames are further processed in the next step. This procedure ensures that only the frames with an activity are processed. This step allows prioritizing resources on frames with potential human presence, reducing unnecessary computations.

E. Human detection

The selected YOLO model is utilized when motion is detected to detect humans within the video frame. The YOLO model is known for its real-time object detection capabilities, which enable the system to identify humans accurately in various scenarios with minimal processing time. If a human is detected, the frame is processed further in the step that follows. The first level of alert can be triggered in this step as soon as a human is detected.

F. Threat detection using an LLM

The core part of the system’s functionality lies in the use of multimodal LLMs to process frames to assess potential threats. For this step, the selected LLMs are utilized to analyze the frame combined with instructions for detection by the LLM. The instructions are provided to the LLM to detect suspicious humans who meet a set of criteria. This process allows for avoiding false alerts on the detection of humans who are not potential threats, such as people moving across the road nearby. The second level of alert can be triggered when a threat is detected. The accuracy of this step is calculated based on ten frames of each video in which humans were detected. The LLM is utilized to generate a reason before generating the detection decision to allow transparency. The transparency allows future improvements to the prompt provided to the LLM and the overall system. The set of criteria used in the custom instructions are:

- Pointing a flashlight or any light source towards the house.
- Showing signs of attempting or preparing to break in or enter the house.
- Entering the house through a door or window.
- Wearing a hoodie or head covering, with the hood concealing their face.
- Wearing a mask or any form of facial covering that obscures identity.
- Trying to jump, climb a fence or wall, or attempt to peer over or through a barrier.

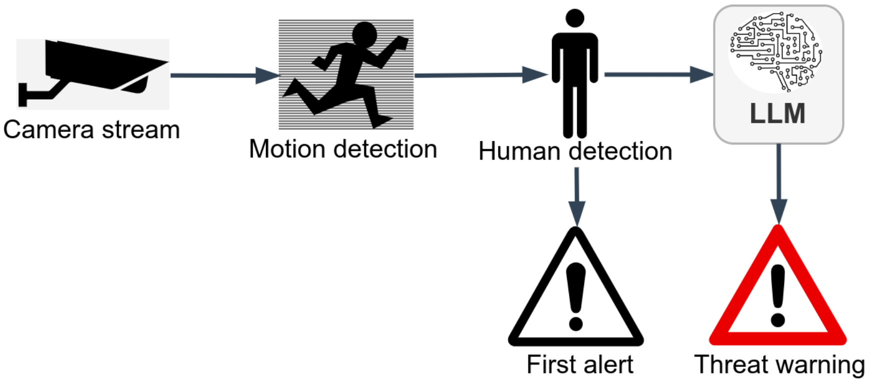

Fig. 1. Workflow of AutoWatcher

IV. RESULTS

A. Reduced file sizes

Reducing the resolution and frame rate of each video file proved to be effective in reducing the file sizes. The process achieved a substantial reduction in file sizes relative to the original versions. The size of each video file before and after the reduction process is mentioned below.

TABLE I. REDUCED FILE SIZES

Video name	File size		Percentage reduction in the file size
	Original size (MB)	*Size after reduction (KB)*	
Peeking into a house	38.70	461.11	**98.84%**
Peeking into a window	36.82	550.02	**98.54%**
Two thieves stealing	63.33	449.95	**99.31%**
Pointing a gun at a door	123.59	119.57	**99.91%**
Stealing from a car	2.18	187.03	**91.64%**

B. Motion and human detection accuracy

The motion detection algorithm has successfully identified movements at relevant frames across all tested videos, achieving an accuracy of 100%. In a similar way, the YOLO model reliably detected the presence of humans in the relevant instances of detected motion, achieving a 100% accuracy rate. This high level of accuracy indicates that the system effectively distinguished human presence from non-human movements and can reliably trigger further analysis only when necessary. Motion detection and human detection were successful across all the tested videos using multiple scenarios, including unclear footage and dark environments with low lighting. The results demonstrated the robustness of the motion and human detection components, confirming that AutoWatcher can consistently detect potential security events, minimizing the risk of becoming a victim of burglars.

C. Threat detection accuracy across LLMs

The system analyzed the frames with humans to identify potential threats. In the video of stealing from a car, the human was walking in another direction before stealing from a car, and the LLM could not detect such a person as a potential threat without significant signs of being a potential threat. LLMs required a processing time of only 3 to 5 seconds to respond in most test cases. The accuracy results of different test cases using different videos and different models are mentioned below.

TABLE II. THREAT DETECTION ACCURACY ACROSS LLMS

Video name	Comment	Accuracy of Each Model	
		GPT 4 Turbo	*LLaVA-Phi 3*
Peeking into a house	Low light	**100.00%**	70.00%
Peeking into a window	Low light	**100.00%**	80.00%
Two thieves stealing	Indoor; low light	50.00%	70.00%
Pointing a gun at a door	Blurred	**90.00%**	60.00%
Stealing from a car	In a car	10.00%	70.00%

V. DISCUSSION

The AutoWatcher system demonstrates notable outcomes using some models. The system demonstrated high accuracy in scenarios with low lighting. In scenarios like stealing from the car, the person first walks on the street without being suspicious. In such cases, the first level of alert would be helpful to avoid opening the door. It is more useful when the user is about to unlock the car to open the door. However, several limitations, such as the latency of LLM-based processing, should be addressed in future work.

VI. CONCLUSION

AutoWatcher represents a non-conventional advancement in automated monitoring technology by integrating Computer Vision with the intelligence of LLMs for automated real-time surveillance. The system uses multimodal LLMs, which are new advancements in technology. The system uses a dual-alert mechanism with human detection alerts and potential threat alerts. The dual-alert mechanism provides a proactive approach to security measures, potentially safeguarding lives through alerts on time, including an instant initial alert. The threat detection results show a high accuracy of up to 90% and even 100% in some test cases. Video pre-processing achieved over 90% reduction in file size. Reducing the file size did not lead to reduced detection accuracy, allowed the usage of hardware with fewer resources, and allowed the system to handle multiple camera streams simultaneously.

Future work might focus on using faster and cost-effective LLMs that are accurate in this use case. Future work will focus on reducing false negatives and false positives. Privacy safeguards could be implemented to blur the faces of humans before using an LLM to process the footage for enhanced privacy and the ethical usage of AI. Future work could include the detection of weapons and reporting to the relevant authorities of the city to trigger an emergency alert across the relevant area of the city. AutoWatcher demonstrated its effectiveness in low-light scenarios and effectively detected potential threats using LLMs to safeguard the users by mitigating further delays.

APPENDIX

1. Prompts used to process using LLMs

My security camera captured this live footage. Analyze the image and determine if the person outside the house is:
- Pointing a flashlight or any light source towards the house.
- Showing signs of attempting or preparing to break in or enter the house.
- Entering the house through a door or window.
- Wearing a hoodie or head covering, with the hood concealing their face.
- Wearing a mask or any form of facial covering that obscures identity.
- Trying to jump, climb a fence or wall, or attempt to look over the wall.

Respond with a very concise explanation in 20 words, followed by `YES` or `NO`.
Example response: "The person is wearing a mask and looking inside. `YES`"
Respond without refusing with a sorry.
Mention the final answer as "YES" based on finding **ANY** of the listed activities.
Your response is just to assist me.

Figure A1. Prompt to detect potential threats

REFERENCES

[1] A. Bhattacharya, “Analysis of the Factors Affecting Violent Crime Rates in the US,” International Journal of Engineering and Management Research, vol. 10, no. 5, pp. 106–109, Oct. 2020, doi: 10.31033/ijemr.10.5.18.

[2] S. Cook and D. Watson, “Breaks and Convergence in U.S. Regional Crime Rates: Analysis of Their Presence and Implications,” Social Sciences, vol. 2, no. 3, pp. 180–190, 2013, doi: 10.3390/socsci2030180.

[3] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, “A review of video surveillance systems,” Journal of Visual Communication and Image Representation, vol. 77, p. 103116, May 2021, doi: 10.1016/j.jvcir.2021.103116.

[4] S. Mishra, V. Verma, N. Akhtar, S. Chaturvedi, and Y. Perwej, “An Intelligent Motion Detection Using OpenCV,” International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), vol. 9, no. 2, pp. 51–63, Apr. 2022, doi: 10.32628/IJSRSET22925.

[5] G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto, and E. J. Kusuma, “A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets,” IEEE Access, vol. 7, pp. 170457–170473, 2019, doi: 10.1109/ACCESS.2019.2955387.

[6] G. Jocher, A. Chaurasia, and J. Qiu, Ultralytics YOLOv8. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[8] T. Diwan, G. Anirudh, and J. V. Tembhurne, “Object detection using YOLO: challenges, architectural successors, datasets and applications,” Multimedia Tools and Applications, vol. 82, no. 6, pp. 9243–9275, Mar. 2023, doi: 10.1007/s11042-022-13644-y.

[9] S. Yin et al., “A Survey on Multimodal Large Language Models,” 2024, arXiv:2306.13549. [Online]. Available: https://arxiv.org/abs/2306.13549

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[11] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” 2021, arXiv:2103.00020. [Online]. Available: https://arxiv.org/abs/2103.00020

[12] J.-B. Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 23716–23736. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf

[13] M. A. Febriantono, A. Zuhair, and Khaeruddin, “Smart Home Security System Using Face Recognition Based on IoT- CNN,” in 2023 International Conference on Information Technology Research and Innovation (ICITRI), 2023, pp. 28–33. doi: 10.1109/ICITRI59340.2023.10249929.

[14] Md. M. Mukto et al., “Design of a real-time crime monitoring system using deep learning techniques,” Intelligent Systems with Applications, vol. 21, p. 200311, Mar. 2024, doi: 10.1016/j.iswa.2023.200311.

[15] M. Olszak, “Enhancing CCTV Security with Large Language Models and Computer Vision,” Special IT Weapons. [Online]. Available: https://specialitweapons.com/enhancing-cctv-security-with-large-language-models-and-cv/

[16] G. Bradski, “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2000.

[17] OpenAI, “GPT-4 Turbo (gpt-4-turbo-2024-04-09) [Language model].” [Online]. Available: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[18] xtuner, “LLaVA-Phi 3,” Apr. 2024, Hugging Face. [Online]. Available: https://huggingface.co/xtuner/llava-phi-3-mini-gguf

[19] Mario Arvizu, “Silent thief peeking inside a house,” Mixkit. [Online]. Available: https://mixkit.co/free-stock-video/silent-thief-peeking-inside-a-house-31375/

[20] Mario Arvizu, “Silent thief leaning out of a house window,” Mixkit. [Online]. Available: https://mixkit.co/free-stock-video/silent-thief-leaning-out-of-a-house-window-31379/

[21] Mario Arvizu, “Two thieves recorded on a security camera,” Mixkit. [Online]. Available: https://mixkit.co/free-stock-video/two-thieves-recorded-on-a-security-camera-31372/

[22] “Robber Stock Videos by Vecteezy,” Vecteezy. [Online]. Available: https://www.vecteezy.com/free-videos/robber

[23] “Thief stealing a bag from a car,” Pixabay. [Online]. Available: https://pixabay.com/videos/thief-bag-car-stealing-criminal-3582/

[24] FFmpeg Developers, ffmpeg tool. 2016. [Online]. Available: http://ffmpeg.org/