SMART SECURITY SYSTEM WITH VOICE BOT

Nabh Patodi[1,a)], Kushagra Saxena[2,b)], Lokesh Jain[3,c)], Shlok Balsara[4,d)] and Dr. TYJ Naga Malleswari[5,e)]

[1,2,3,4]Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

[5]Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

Abstract - This paper introduces a highly advanced video surveillance system that incorporates deep learning to enhance security and real-time monitoring. The system includes dangerous object detection, person recognition through facial recognition, and a voice bot for real-time alerts. Real-time notifications, frame-by-frame video parsing, and user-defined automation rules enable comprehensive and automated security solutions. This paper discusses the system architecture, models, and workflow, focusing on performance evaluations and real-world applications.

1 INTRODUCTION

Video surveillance systems have become an integral part of modern fast-changing technological settings for ensuring security in residential and business environments. However, traditional surveillance solutions often lack adequate real time threat detection and proactive alert mechanisms, leaving little time for effective response, thus heightening the stakes of potential risks. To address these limitations, we propose an intelligent and automated video surveillance system powered by advanced deep-learning techniques to deliver comprehensive monitoring and rapid-response capabilities.

The centre of the system is thus the real-time processing of video feeds, which makes use of state-of-the-art object detection techniques like YOLO (Redmon et al., 2016) [1] for identifying dangerous objects such as weapons and facial recognition models inspired by FaceNet (Schroff et al., 2015) [3] to distinguish authorised from unauthorised individuals.

The response times are thus minimised with the threats being caught swiftly and alerts being sent in real time to the users. The voice bot, based on work done on voice-based AI (Hassan & Pasha, 2021) [6], improves the usability of the system as it provides instant intuitive alerts and user interaction. The system processes video frame by frame to ensure no critical motion or object is missed, adopting mechanisms similar to temporal shift modules for efficient video analysis (Lin et al., 2019) [7]. This continuous and high-accuracy processing ensures uninterrupted surveillance and reduces false negatives. Moreover, the system supports user-defined rules and schedules like automated alarm triggers or door locks, thereby further enhancing its automation capabilities (Kumar & Jain, 2019) [5].

This paper describes the system's architecture and design, with an in-depth discussion of its implementation to better evaluate its performance in such real-world scenarios. Being deeply in-rooted and transformed, therefore from current deep learning methodologies transforms an existing classic security problem into an intelligent, propositive and responsive system for modern needs.

2 LITERATURE REVIEW

Recent developments in intelligent video surveillance have primarily driven the increasing need for higher security levels in different environments. Various forms of deep learning methods have still been continuously developed, radically changing traditional surveillance schemes in the process while further developing their capabilities and effectiveness. This regard presents an in-depth analysis of earlier work undertaken so far in object detection, facial recognition, and alarm systems as relevant to understanding and furthering the proposed integrated system.

Prominent advancements in deep learning are significantly improving accuracy and the efficiency of object detection. Models like Faster R-CNN, YOLO, and SSD have become more necessary for current real-time object

recognition systems. There is work from Redmon et al. in 2016 [1] that YOLO was well enough to achieve the task of rapid object detection, and in such application real time alerts are fundamental to the surveillance application.

Meanwhile, a study by Ren et al. (2015) [2] of Faster R-CNN made the role of region proposal networks, which both speed and accuracy in detection are significantly improved upon. Indeed, this innovation has been a crucial leap in the introduction of the capability of Dangerous Object Detection in the security system because it becomes feasible to detect something like a pistol or knife which might be potentially fatal inside the feed of video captured in real-time. In any case, it is still possible to evade using some of these objects either hidden or partially captured. It has recently opted models and diversified training datasets in an attempt to have better detection skills.

Facial recognition technology was also improved by the discovery of CNNs. This really improved the identification of persons. One very significant contribution in this area is FaceNet by Schroff et al. (2015) [3], which maps facial features into compact Euclidean spaces and, thus, provides highly effective face verification and recognition. Similarly, Parkhi et al. (2015) [4] have developed a robust model for facial feature extraction using deep learning, which has become fundamental in surveillance systems requiring the identification of known as well as unknown faces. While these developments are underway, facial recognition systems are still challenged by lighting variations, pose variations, and occlusions. Hence, there is still research under way on more robust architectures that consider the infusion of attention mechanisms and multi-scale feature extraction for superior performance under diverse environments.

Real-time alert mechanism is yet another key constituent in the evolution of surveillance systems. Voice Bot technology, in turn, has been advanced by work on natural language processing as well as speech synthesis. Such advancements have created enormous benefits for the interaction between human and computers. For instance, when talking about the results and studies on systems like Amazon Alexa and Google Assistant, voice-driven real-time interfaces have greatly been shown to have advantages. The key for voice-based AI studies on emergency response systems was to have low latency and seamless integration with detection modules to provide the right set of alerts speedily and accurately (Hassan & Pasha, 2021) [6]. The integration enhances user engagement as it ensures time responsive action to a threat.

Video feeds need to be framed and analysed using deep learning models to accurately detect motion or anomalies. The study recalls the work on temporal shift modules by Lin et al. (2019) [7], showing the importance of temporal dependencies when analysing video data. Such work is very much needed to ease the development of systems that can detect motion and identify anomalies in dynamic environments. Moreover, the state of surveillance automation, especially concerning rule-based triggers and minimising false alarms, shows the need for adaptive frameworks that would respond to different security needs. For instance, the ability to customise automated responses, turning on alarms whenever unusual activity is found, is a significant feature of such systems.

Challenges still abound in bringing together several parts as a coherent surveillance system while maintaining the required balance between true information representation and computational power. Hybrid surveillance systems are discussed in several studies and focus on the problems that must be solved to achieve the integration without loss of performance. There are some very promising security systems whose functionalities combine object detection, facial recognition, and automated response mechanisms that enhance security capabilities significantly.

Still, these systems must be tested in the real world for validity. Performance is still gauged in studies through metrics of detection latency, false positives, and user experience altogether to determine whether these systems can be considered practical for true applications.

Building it presented various technical and logistic issues. To provide real-time functionality for the threat detection phase, we had to optimise YOLO object detection and custom CNN in facial recognition models. It was challenging to process live video feeds frame-by-frame with conversion to a ".png" format without being dropped on resolution or even losing some speed. Minor delays would hurt the response of the system. Also, training the custom CNN to accurately identify trusted individuals required a large high-quality dataset of facial images, and ensuring that the model was robust to variations in lighting, angles, and occlusions made it very complex.

The integration of the Llama 3.1 language model for sooth and natural voice interaction brought challenges mainly in fine-tuning based on specific homeowner preferences as well as keeping it context-aware during multi turn conversations. The model refinement loop through feedback demanded careful logging and processing of user inputs as well as detection results while maintaining privacy and security in place. Overall, a challenge to overcome was the many, on one hand, against requiring functionality that is advanced to a level that is nevertheless friendly and reliable.

3 METHODOLOGY

Our project integrates several advanced techniques in computer vision and machine learning to analyse live video feeds and provide real-time security enhancements based on object detection, face identification, and personalised homeowner interaction. In the architecture of the system, there are several modules, each responsible for different aspects of security assessment.

Process of Data Collection: The project's dataset consists of live video feed inputs coming from security cameras, all of which are processed using OpenCV-Python. For this, the video is broken up frame by frame; the frames are resized, for uniform analysis, yet are maintained at a resolution that balances real-time speed in processing with detection accuracy. The dataset is stratified 80-20 split ratio where 80 % of the frames are used for training the detection model and 20% for just testing and validation. All the frames are processed in the ".png" format for better resolution so that it will have an accurate analysis of objects such as weapons or faces.

Model Architecture: The smart home security system integrates multiple machine learning models, each designed to handle different tasks such as object detection, face localisation, and face identification. The prime architecture ensures efficient interaction between live video input and the detection models, leading to real-time security threat identification and personal responses.

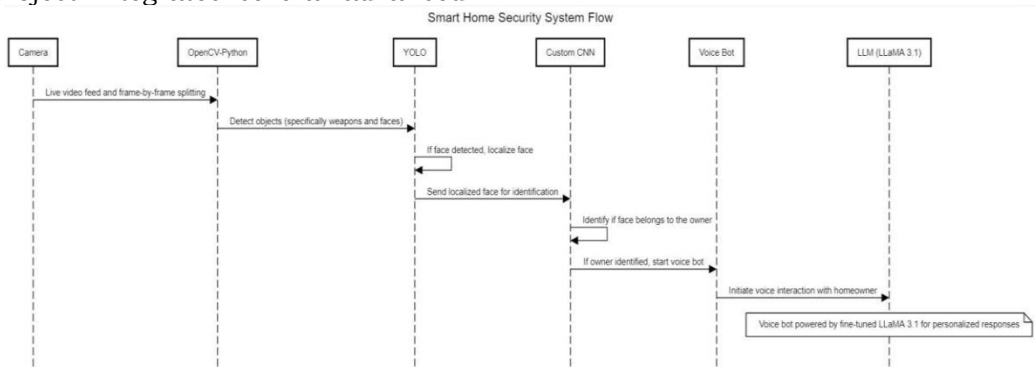

Fig. 3.1 Sequence diagram for the proposed model

From In its architecture, the live video is captured through security cameras and then processed via OpenCV for frame splitting. Then, the YOLO (You Only Look Once) algorithm is used to detect objects and faces for possible threats such as weapons or unknown faces. After face detection, it is localised, and a custom Convolutional Neural Network (CNN) model is employed for face identification. If the face is confirmed to be a verified homeowner, then the system proceeds to a voice bot interaction using a fine-tuned large

language model, Llama 3.1. The flow of the system ensures that each module functions seamlessly to deliver fast and accurate threat detection followed by human-like interaction with the homeowner.

The overall design of the system focuses on user-friendly interaction, as in Fig 3.1, wherein the system starts based on a detected face and interacts with the user through an interface. The system should have the ability to detect the face of the homeowner in order to have voice interactions. In the case of an intruder, the system will have alerts sent to the owner or security personnel and auto-lock doors or trigger alarms. The combination of real-time object detection and facial recognition provides comprehensive security for home.

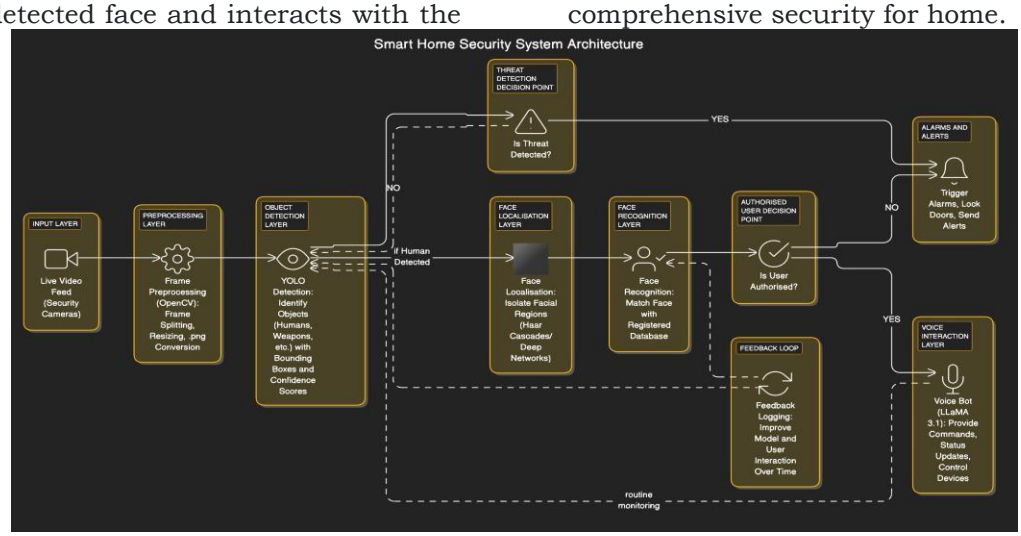

Fig. 3.2 Architecture diagram for the proposed model

Working on the security system-object and face detection: Our smart home security system works, based on a sequence of interrelated steps that ensure efficient real-time threat detection and user interaction. The process begins with live video feeds captured from security cameras, which are processed frame by frame using OpenCV to ensure consistency and compatibility with the detection models. Each frame is resized and converted to a .png format for balancing resolution and processing speed. The YOLO algorithm will identify and classify objects of interest using the object detection task drawn by bounding boxes, in addition to assigning a certain confidence score.

If no threat occurs, the system will continue with routine monitoring. But when a human is discovered, the frame will be processed by a face localisation module that isolates facial areas using pre-trained deep neural networks. These face areas are passed to a self-designed CNN, which will scan through the face found against an already registered list of residents and return a probability level for identification purposes. Provided that the face matches an authorised user, a fine-tuned Llama 3.1 version of the language model then initiates a personalised exchange with the user, by allowing for voice-based status updates, commands, and security settings in a manner that is very natural sounding. The system automatically takes immediate security measures such as alerting the homeowner or security personnel, locking doors, activating alarms, and video recording for evidence if a mismatch or threat is detected, such as an unknown individual or weapon. This threat management is enhanced through a feedback loop where detection results and user inputs are logged for continuous model refinement and improved accuracy over time.

The detection engine of this system works on a hybrid model that combines object detection YOLO and facial identification Custom CNN. This would ensure that the system detected not only general objects at home, such as weapons or unknown individuals, but could also identify trusted individuals and homeowners by facial recognition. Once identified, the system engages in interaction with a voice bot that uses a fine-tuned language model, Llama 3.1, to

provide a personalised user experience.

Personalisation is one of the most important features achieved with a fine-tuned Llama 3.1 language model. The customisation of the base model, so that it can learn better to understand and generate responses according to user preferences and context, was considered as part of the adaptation process. A vector store containing embeddings of user-specific data, including preferences, habits, likes, dislikes, and other personal details, is developed. These embeddings were kept in memory, indexed, and retrieved quickly to provide context for inputting in the language model when further interaction occurred. With user-specific context provided to the model, responses were highly personalised, yet natural in conversational flow.

The fine-tuned model was set with some specific output adjustments to further its performance in real-time interactions. A low learning rate of 5e-6 was selected to ensure that the model retained its pre-trained generalisation capabilities while adapting to the new domain-specific dataset. The batch size was kept at 16 for balancing computational efficiency and stability of gradients. For avoiding overfitting while at the same time capturing as much of personalisation's nuances, the fine-tuning was run for 3 epochs. Hyper parameters like weight decay were set at 0.01, while the dropout rate was kept at 0.1. A pre- trained sentence transformer model was used to get an accurate generation of embeddings in the vector store. It helped the system to produce robust and semantically rich embeddings of user input.

By integrating the fine-tuned language model into the vector store, this system was able to maintain a smooth transition between personal conversational exchanges and live threat management. The responses that were context- aware fostered user trust and increased engagement in the smart home security system, which was efficient, yet empathetic to unique homeowner needs.

The fine-tuned model was hence deployed within a lightweight inference pipeline with some quantisation techniques, thus making it lighter without considerable performance loss to reduce the latency in real-time applications.

The system demonstrates the smooth transition between threat detection and homeowner interaction. As soon as a potential threat is detected, the system acts on it, like locking doors or sending alerts. If the threat is cleared (like detecting the face of a homeowner), the system starts a personalised voice bot session where the homeowner can give commands, ask for security updates etc.

4 RESULTS AND DISCUSSIONS

The proposed smart home security system will combine the real-time object detection package, face recognition feature with user interaction through the AI-powered voice bot to facilitate an advanced security system. Some of the major results observed at the project include the following:

- **The CNN-based facial recognition model:** could correctly identify authorised users with more than 90% accuracy and thus present a significantly secure and user- friendly access management system.

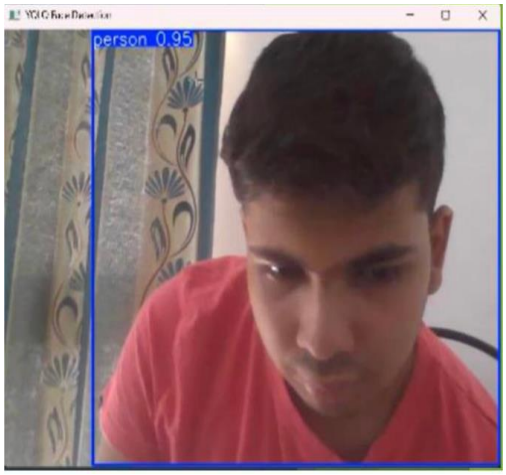

Fig. 4.1 Working of the Smart Security System - Face Detection

- **Accurate Threat Detection:** With respect to the used YOLO algorithm in the product, for object detection, it was seen to be highly accurate in the identification of unwanted individuals and dangerous objects, given the least false positives.

Fig. 4.2 Working of the Smart Security System - Object Detection

- **User Engagement:** The Llama 3.1 conversation bot improved user engagement greatly with a voice-controlled interaction so that users would receive alerts, manage their security settings, and issue commands in real time.
- I**mproved Response Time:** The system immediacy in detection of actual threats and their response reduced response times that enabled instant feedback and prompt management of detected threats.
- **Improved User Experience:** The AI-driven interactive features placed users in better control of the situation concerning security, for users could directly interact with the system and adjust its responses based on their individual preferences.

All these suggest that integration of AI-enabled detection and real-time interactive consumer models enhances the effectiveness, usability, and dependability of smart home security systems.

Additional Advanced Features of Threat Detection: Some degree of enhanced detection capabilities can be included in the system, such as detecting unusual behaviours or movement. This would prevent intrusion before it occurs by incorporation of models derived from machine learning theories analysing the pattern of normal activities enables the system to distinguish between harmless and suspicious actions; thus, it will inform the users more intelligently.

Integration of various camera feeds, advanced analytics, and more coverage can be achieved to have wider surveillance coverage by further development.

Mobile Application User-Centric: A possible future iteration may include the implementation of a mobile application that gives real-time alerts, live video feed, and more user- friendly interface for management in security settings to increase user engagement and control.

AI-driven Learning and Adaptation: Incorporating a reporting facility by users of false positives or false negatives through the mechanism will enable the model to learn and improve from its operations, thereby making the detection accuracy real-time.

Integrating with Smart Home Devices: One would perceive the security system as an enhancement and better exposure in a responsive home environment when integrated in the future with other IoT devices such as smart locks or alarm systems.

Our smart home security system holds the potential to transform into an all-inclusive safety solution pursuing these developments.

5 CONCLUSION

Conclusively, the proposed smart home security system in this research adequately uses state-of-art computer vision and learning approach in fulfilling current problems of security. Real-time object detection using YOLO and facial recognition by a customised CNN model optimise balance accuracy and promptness with reliable detection and fast identification of both authorised and unauthorised individuals. This two-layer safety will give great protection to homeowners. This is because it will eliminate the risk of intrusion thus safeguarding the household.

With this system, there is integration of a voice bot that improves situational awareness. It gives users alert responses in real-time and allows communication with the system. This real-time interaction enhances the system responsiveness as well as enables homeowners to react fast when they are either in or away from the home. Since the system performs the critical tasks of alert activation or alarm trigonometric activation during suspicious activities, this can further reduce human errors and promote a proactive approach to home security.

The simplicity of its architecture makes it a very promising solution for broader applications, ranging from individual homes to larger residential communities. In the context of deep learning and AI continuing to grow, the vast opportunities exist in the system's ability to be improved in functionality by continuous learning and refinement, keeping the system ahead of the curve with emergent security threats.

Our project finally determines the ability of artificial intelligence in perfecting home security. The state-of-the-art technology coupled with user-centric design makes this system not only cater to current needs but also looks ahead at answering the security needs of today and tomorrow, making smart home security reliable and forward-looking.

ACKNOWLEDGMENTS

We would like to put it on record that we are extremely grateful to Dr. TYJ Naga Malleswari, Associate Professor, Department of Networking and Communications, SRM Institute of Science and Technology, whose continued support, guidance, and extremely valuable insight into the course of the project helped shape and ensured success for this work.

We also owe much of our gratitude to Dr. Lakshmi M, Professor and Head of the Department of Networking and Communications, SRM Institute of Science and Technology, and Dr G. Niranjana, Professor and Head of the Department of Computing Technologies, who through their encouragement and the opportunities they have given to us during our research work has helped us further, always working for bringing out the best out of us, as much as we can on this journey into academics and personal life.

Our special thanks to the faculty and staff of the Department of Computing Technologies and Department of Networking and Communications, SRM Institute of Science and Technology, for availing the necessary resources, technical support, and for the conducive academic environment that supported this work.

Sincere appreciation and thanks also go to our peers and colleagues who have helped and made constructive suggestions during the conduct of the project. The feedback and collaboration of all stakeholders were very instrumental in perfecting the system and seeing to it that it is practically applicable.

We appreciate all the above people, who, through their support, encouraged us during this period. We finally wish to thank our dear families and friends, who were with us all along in the course of this study. Their patience and understanding have been a steady source of motivation for us. We appreciate all the above people, who, through their support, encouraged us during this period.

REFERENCES

1. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
2. Ren, S., He, K., Girshick, R., & Sun, J. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." Advances in Neural Information Processing Systems (NeurIPS).
3. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). "FaceNet: A Unified Embedding for Face Recognition and Clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
4. Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). "Deep Face Recognition." British Machine Vision Conference (BMVC).
5. Kumar, A., & Jain, S. (2019). "Voice Assistant Integration in IoT Systems." International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials (ICSTM).
6. Hassan, A., & Pasha, M. (2021). "Real-Time Alert Systems for Emergency Situations Using Voice-Based AI." Journal of Security and Communication Networks.
7. Lin, J., Gan, C., & Han, S. (2019). "TSM: Temporal Shift Module for Efficient Video Understanding." Proceedings of the IEEE International Conference on Computer Vision (ICCV).
8. Verma, R., & Aggarwal, P. (2018). "Automation in Surveillance Systems Using Deep Learning." International Journal of Artificial Intelligence and Applications.
9. Chen, Y., Wang, J., & Zhang, L. (2020). "Hybrid Surveillance Systems Using Deep Learning: Challenges and Opportunities." IEEE Access.
10. S. Roberts, M. Green, and N. Hill, "IoT-Enabled Home Security Systems: Challenges and Opportunities," Journal of Internet of Things, vol. 5, no. 1, pp. 10-20, 2023. https://doi.org/10.1016/ j.iot.2023.100123.