IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

SafeSight: Privacy-Preserving AI Passive Monitoring System for Situational and Health Awareness

Mridula Murugan

South Brunswick High School, 750 Ridge Rd, Monmouth Junction, NJ 08852, USA; mridula.murugan@gmail.com

ABSTRACT: Real-time monitoring systems are being increasingly utilized as an option to protect vulnerable populations, but current solutions heavily depend on sensors that reduce effectiveness and create privacy concerns. The proposed SafeSight system addresses this gap by applying an array of models to perform scene understanding, object detection, action recognition, and motion anomaly detection to enable contextual privacy masking. We evaluated SafeSight with benchmark datasets (UCF50, GMDCSA) and live video feeds. It achieves up to 0.867 F1 score for sedentary activity recognition and over 90% scene context accuracy within household environments. The use of artificial intelligence through deep learning and vision language models revolutionizes live video analysis to make accurate decisions with temporal scene analysis, static and dynamic events, and privacy-protecting contoured filters for sensitive locations. SafeSight could feasibly expedite alerts while monitoring high-risk situations where medical attention is required immediately, eventually being scaled to larger scenarios for healthcare, education, and public safety.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Health Monitoring, Vision Language Model, Ultralytics.

Introduction

Real-time video analysis systems are becoming an increasingly valuable tool for monitoring the safety and well-being of vulnerable populations, including senior citizens, unattended children, and individuals with health conditions. For elderly individuals living alone at home or in assisted care facilities, these systems are critical for detecting potentially fatal incidents and alerting the caregivers.[1] Similarly, young children at home or in childcare settings benefit from continuous supervision to prevent and respond to accidents and emergencies. Beyond individual care, real-time monitoring has applications in law enforcement and institutional settings, such as detecting disturbances in prisons or ensuring safety in public areas. These systems offer situational awareness, enabling caregivers, family members, or authorized personnel to remotely track current activity within the monitored property.

Timely intervention during medical emergencies, such as falls, seizures, or strokes, can significantly improve survival and recovery outcomes.[2] In the United States alone, over 14 million people, primarily seniors over 65, experience falls annually. Similarly, strokes impact approximately 795,000 individuals each year, and rapid intervention within the first hour is critical to reducing complications. Seizures lasting longer than five minutes also require immediate medical attention. Effective monitoring systems must be capable of detecting both static events, such as lying down or sitting, and dynamic events, such as walking or suddenly collapsing, to enable rapid response.

There are various challenges associated with real-time monitoring systems. They raise serious concerns about preserving the user's privacy, especially in sensitive locations of the property, like bedrooms or bathrooms, where supervision of vulnerable populations is still required. Another challenge is that these systems often over-trigger alerts for non-emergency events or fail to raise on-time alerts for genuine emergencies. Existing systems often use a mix of wearable and infrastructure-based sensors.[3] Wearables such as accelerometers and heart rate monitors can provide continuous physiological data,[4] but their reliability is limited by battery constraints, inconsistent use, environmental sensitivity, and user discomfort or forgetfulness.[5,6] Infrastructure-based systems often require a high computational load, delaying detection and reducing the overall effectiveness of interventions.

To overcome these challenges and address user needs, real-time alerting systems must maintain a balance between responsiveness and user privacy. Monitoring should not result in a sense of constant surveillance, and alerts should only be triggered when necessary, as long as any privacy measures are taken.

Infrastructure-based sensors, such as cameras, depth sensors, radar, and LiDAR, offer more consistent and non-intrusive monitoring by capturing visual and spatial data without requiring the subject to wear any devices. In the SafeSight project, we adopt camera-based infrastructure sensors to obtain detailed visual input and contextual awareness across monitored spaces. Computer vision techniques, including object detection, tracking, pose estimation, and activity recognition, form the foundation for analyzing real-time video streams. By leveraging deep learning frameworks such as TensorFlow,[7] PyTorch,[8] YOLO,[9] and OpenCV,[10] SafeSight builds an accurate and scalable monitoring system.

To further enhance performance and contextual understanding, SafeSight incorporates Vision-Language Models (VLMs) using tools such as Olama. Unlike traditional deep learning models, VLMs are capable of reasoning across both visual and

DOI: 10.36838/v8i1.98

textual data. They can generalize to a broad range of objects and actions beyond those seen during training and perform complex tasks like visual question answering and scene interpretation. This integration enables SafeSight to deliver a context-aware, privacy-conscious, and responsive solution for real-time safety monitoring. To achieve this, we incorporate several key contributions:

Scene-aware privacy-preserving real-time alert generation: We developed algorithms to identify and alert on unusual events while minimizing the exposure of sensitive information.

LLM-driven rules engine: We utilized large language models to define and enforce rules about valid and invalid behaviors within a given scene.

Real-time assessment of mobility state: We developed techniques to assess an individual's mobility state independently of their current posture, enabling the detection of subtle changes in activity levels.

Related Works:

Within previous studies regarding real-time monitoring systems for vulnerable populations, wearable sensors have been researched for their ability to detect falls and monitor vitals. For instance, fall detection systems that use accelerometers exhibited high accuracy in controlled environments. However, in real-world scenarios, there are other factors, such as user mobility and battery life, that could limit success.[11] Additionally, wearable devices such as heart rate monitors and smartwatches can provide continuous health data, but become less dependable due to the user's forgetfulness and improper placement.[12] Although there are significant advancements in wearable technology, the continued limitations stress the need for solutions that rely on infrastructure-based monitoring instead of wearable sensors. Systems combining wearable sensors with environmental context have shown improved accuracy and reliability in detecting anomalies.[3]

Low-resolution infrared arrays with 3D convolutional neural networks achieve accurate privacy-preserving fall detection, but struggle in cluttered environments.[13] Optical elements may reduce identifiable data yet compromise contextual details.[14] Privacy-preserving cameras and neuromorphic sensors limit visual information, but face challenges in dynamic scenes.[15-16]

Although depth sensors, radar sensors, or LiDAR sensors can be used to create detailed 3D maps of environments, these technologies are typically expensive and complex to implement on a larger scale.[17] By using camera-based systems and computer vision techniques, many of these limitations can be reduced or eliminated. Frameworks for deep learning, such as TensorFlow, MediaPipe,[18] and YOLO, enable applications related to human activity recognition and anomalous event detection. When these techniques are used together with depth data, they create strong detection systems in obstructed environments.[19] Vision-language models (VLMs) have recently been used for real-time understanding using camera-based systems. The coordination between visual and textual data allows for a more sophisticated understanding and reasoning of the conditions being monitored.[20] The advantage of these models is that they use joint representations of images and texts, allowing for increased detection of anomalies and behavior analysis. Empirical works assert that VLMs indeed enhance performance in important tasks like visual question answering and scene interpretation, both of which are crucial for real-time monitoring.[21]

In addition to detection accuracy, privacy preservation is a critical focus in the field. Some existing methods include edge processing and encrypted data transmission, which can protect user privacy while effectively running the system.[22] The SafeSight project uses VLMs, deep learning methods, and appropriate privacy-preserving methods to integrate a reliable monitoring system that effectively addresses the drawbacks of existing solutions.

Existing monitoring approaches face trade-offs between accuracy, cost, and privacy. While wearable sensors capture physiological data, they suffer from compliance and battery issues, whereas infrastructure-based options such as LiDAR and radar are expensive and difficult to scale. Camera-based deep learning approaches improve activity recognition, but often lack semantic reasoning. SafeSight addresses these gaps by combining camera infrastructure with VLM-driven scene understanding, LLM-based rule enforcement, and deep learning techniques. This integrated design enables accurate, privacy-preserving monitoring of vulnerable populations, distinguishing SafeSight as a novel and practical real-time solution.

Methods

Architecture:

The SafeSight system architecture, as seen in Figure 1, is designed to provide comprehensive, real-time monitoring in places like homes, childcare centers, assisted living facilities, etc. This system is centered around the cameras that are strategically placed around the monitoring facility and connected to a pre-processor module. The classifier consists of six modules:

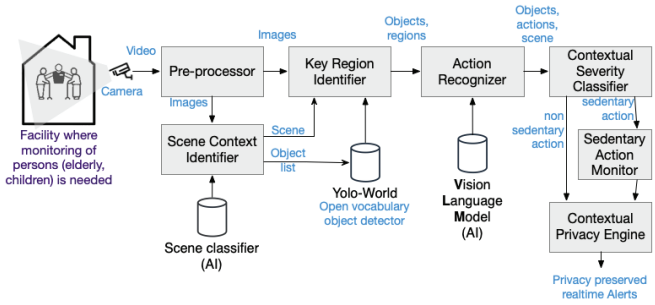

Figure 1: SafeSight system architecture. The system takes the video footage from the real world, extracts the images, and takes them through a series of steps to clean up the images, extract key regions of human interactions, and identify the performed actions of individuals within those regions. After classifying the actions, an informative alert will be triggered along with privacy-protected alert images.

Image Pre-processing module: This module performs frame extraction, color space conversion, and resizing. This step removes any unwanted noise or background artifacts, ensuring consistent, high-quality input into the system's other modules.

Scene Context Identifier: This module can differentiate certain contextual details, such as the room type, the objects that are present, and allowed or safe actions. It utilizes the LLAVA

Moondream2 VLM combined with a large language model (LLM) to describe the scene in detail.23,24 Once the context of the scene is established, YOLO-World identifies and localizes key objects within the frame. This process generates bounding boxes for recognized objects.

Key Region Identifier: This module segments these into key interaction regions that focus on interactions between individuals and nearby objects or other people. This allows the system to monitor specific scenarios, such as a person cooking with a stove in a kitchen or brushing their teeth in a bathroom.

Action Recognizer: This module uses a Vision Language Module to identify the actions being performed by the individuals within these key regions.

Contextual Severity Classifier: This module classifies these actions into non-sedentary or sedentary categories. Sedentary actions are further scrutinized by the Sedentary Action Monitor to detect joint movements and assess the mobility of the individual. If minimal to no movement is observed within a predefined time frame, an alert is sent out to emergency contacts and services.

Contextual Privacy Engine: This module blurs sensitive regions while preserving the overall shape of the individual's figure. It ensures that alerts are informative while preserving the privacy of the monitored individual, especially in sensitive areas such as bathrooms or bedrooms.

Sedentary Action Monitor:

The SafeSight system follows a classification technique that can effectively categorize human activity in varying scenarios. The large language model (OpenAI GPT-4o) is utilized offline to categorize valid sedentary and non-sedentary actions in different scene contexts (Table 1). Sedentary actions involve limited movement and are mostly related to rest, such as sitting, lying down, or reclining. Conversely, non-sedentary actions involve dynamic movements and are done with significantly more physical exertion, such as walking, standing, cooking, or cleaning. By distinguishing between these action classes, the SafeSight platform can respond to immobility or contextually unusual behavior that may indicate a health emergency.

Table 1: Defines the classification of sedentary and non-sedentary actions across general movement, household activities, and leisure tasks. Sedentary actions encompass low-movement activities like sitting and reclining, while non-sedentary actions involve dynamic tasks such as walking, standing, or cleaning. Accurate distinction aids SafeSight in the timely detection of immobility or unusual behavior.

Activity types	Sedentary actions	Non-sedentary actions
General movement	Sitting	Walking, running
	Lying down	Standing up
	Reclining on couch	Stretching
Household activity	Watching TV	Cooking, cleaning, sweeping, dusting
	Reading	Folding, washing, loading clothes
Leisure	Drawing, painting	Playing ball
	Playing light instruments	Gardening

The ability to accurately classify actions into sedentary and non-sedentary categories improves the real-time monitoring capabilities of the system. For instance, sedentary actions may require closer observation in order to detect immobilization and health concerns, such as a fall or stroke. On the contrary, non-sedentary actions can show that the individual is actively engaged in their current tasks. This framework supports the SafeSight system's situational awareness and its goal of protecting vulnerable populations. Table 1 exemplifies common household activities that fit within both action classes, demonstrating the applications of SafeSight in real-world environments.

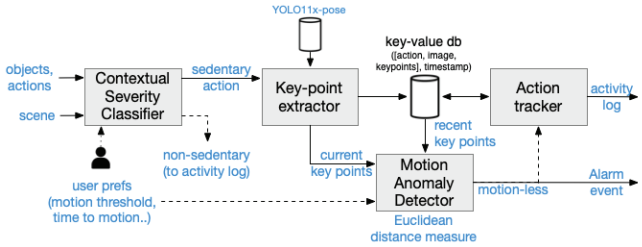

Figure 2: Architecture of the Sedentary Action Monitor, highlighting contextual severity classification of detected actions, pose key-point extraction for monitoring sedentary behaviors, and motion anomaly detection via Euclidean distance calculations. This multi-stage process ensures timely intervention during anomalous events while minimizing false alarms.

The architecture of the Sedentary Action Monitor is shown in Figure 2, portraying the process for analyzing and responding to sedentary behaviors. The first step involves the Contextual Severity Classifier, which evaluates actions detected by the Action Recognizer. As outlined in Table 1, it will then categorize the actions into sedentary or non-sedentary classes based on factors such as the objects interacted with, the subject's location, scene description, and user preferences. The classifier follows specific rules, basing its operations on expert knowledge, user settings, and LLM-generated rules that define valid actions for specific contexts, locations, and situations. Depending on the user's privacy settings, SafeSight will log non-sedentary actions for documentation, but will continue monitoring sedentary actions for any anomalies. This approach ensures that the actions are properly classified and that appropriate steps can be taken. Customizable user preferences enable users to specify mask settings, data access, and alert recipient permissions, ensuring only authorized caregivers receive notifications. While adults and other elderly individuals can provide their consent, the usage of SafeSight for children may require consent from parents or guardians. This approach upholds the user's autonomy and minimizes surveillance concerns, aligning with ethical standards for privacy-preserving AI monitoring.

When sedentary activity has been detected, the Key-point Extractor extracts body key points and stores them in a key-value database along with the corresponding timestamp, image, and action metadata. The Motion Anomaly Detector is continuously fed these key points, allowing it to compare the current key point set with previous ones using a Euclidean distance calculation. If the cumulative motion distance D, which is calculated across a time window T, reaches below

DOI: 10.36838/v8i1.98

a threshold (τ), the system sends an alert, signaling that the subject is motionless and may be in need of intervention. SafeSight's multi-stage architecture ensures accurate monitoring of sedentary behavior, enabling timely detection of potential emergencies while minimizing false alarms. The algorithm incorporated in the Motion Anomaly Detector is as follows:

$P_t = [p_1, p_2, \ldots p_n]$ (key points for current time)
$P_t - k = [p'_1, p'_2, \ldots p'_n]$ (key points for past time)

Euclidean distance = d_k

$$d_k = \sqrt{\sum_{i=1}^{n} (p_i - p_i')^2}$$

Cumulative distance (over time window, T) = $\boldsymbol{D}$

$$D = \sum_{k=1}^{T} d_k = \sum_{k=1}^{T} \sqrt{\sum_{i=1}^{n} (p_i - p_i')^2}$$

Threshold for motion anomaly = $\boldsymbol{\tau}$
If $\boldsymbol{D} < \boldsymbol{\tau}$, motion anomaly detected, an alert is triggered.

Contextual Privacy Engine:

The architecture of the Contextual Privacy Engine, illustrated in Figure 3, plays a role in ensuring privacy during sedentary action monitoring. The initial step involves the Privacy Filter block, which assesses whether privacy masking is necessary based on the scene context, user preferences, and the current action. If the Privacy Filter decides that privacy masking is not required, the alert is dispatched directly, but if it is necessary, the system proceeds to the masking sequence.

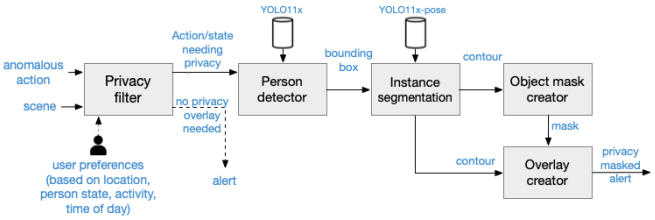

Figure 3: Architecture of the Contextual Privacy Engine, illustrating privacy decision-making based on context, user preferences, precise localization and segmentation of individuals, and application of tailored visual privacy masks. This ensures alerts maintain privacy standards during anomalous events.

The Person Detector localizes the individual precisely in the relevant image region and generates bounding box coordinates. The Instance Segmentation module then uses these coordinates to segment the person within this bounding box and produce a contour for targeted masking. Using the contours, the Object Mask Creator generates a visual privacy mask tailored to the individual, which is then combined with the original image, ensuring both visual integrity and privacy enforcement. Figure 4 shows the original image, followed by the privacy mask and the contextual mask according to the figure's surroundings (bathroom). This process results in a privacy-protected image that will protect user privacy within any alerts shared to necessary contacts. Unlike IR or thermal cameras, which struggle with visual accuracy in low-contrast environments, SafeSight's contour-based masking preserves privacy while retaining contextual awareness and attention to detail.

Figure 4: Sample output of the Contextual Privacy Engine demonstrating localization and segmentation of an individual who is under distress in a bathroom, followed by the creation and application of a tailored visual privacy mask. Original image of immobilized person (left), contour mask of immobilized person (middle), privacy-preserving contour mask overlaid on original image (right).

Implementation:

The SafeSight system was implemented using Python 3.11, leveraging its extensive ecosystem for computer vision and AI-based tasks. Real-time video input is captured using a Microsoft Lifecam HD-3000 webcam. Frames are extracted at a configurable frame rate using OpenCV's video capture and decoding utilities, and subsequently pre-processed (e.g., format conversion, resizing, noise reduction) for analysis.

Scene context is established using the Moondream 2 VQA model, which processes each frame with scene-specific prompts to classify the environment as kitchen, bathroom, living room, etc. To streamline object detection, an offline process using OpenAI's GPT-4o generates scene-specific object lists, which are cached and indexed for fast access. These lists serve as constraints for the Open Vocabulary Object Detector, implemented using the YOLO-World model,[25] ensuring that only context-relevant objects are detected in each scene.

For activity analysis, the same Moondream 2 model is used to infer the action being performed by the subject in the frame. If a sedentary action is detected, the Key-Point Extractor Module uses the YOLO11x-pose model to estimate human pose keypoints. These key points are used to compute motion trends over time for anomaly detection.

To preserve privacy in sensitive contexts, SafeSight incorporates a privacy masking module. First, the subject is localized using the YOLO11x detector. Then, YOLO11x-pose refines this region to produce a segmented body contour.[26] OpenCV functions, such as cv2.fillPoly and cv2.addWeighted, are used to generate and blend a privacy mask onto the frame. The resulting privacy-protected image is used in any alerts, ensuring that context is preserved while respecting the subject's privacy preferences. This modular implementation allows SafeSight to balance accuracy, responsiveness, and privacy protection in real-world monitoring scenarios.

System Setup:

The SafeSight system was implemented using an Intel i9-based NUC platform that can support real-time monitoring and the processing requirements of the application. The system features an Intel Core i9-9980 HK processor with 8 cores, 16 threads, and a base clock speed of 3.4 GHz, with 64 GB DDR4 RAM. For graphical processing, it incorporates an NVIDIA®

GeForce® RTX 3060 GPU with 12 GB of memory. The platform operates on Ubuntu 22.04 LTS and is tested using an external Microsoft Lifecam HD-3000 webcam.

Datasets:

To evaluate SafeSight's accuracy in action recognition and fall detection, benchmark datasets and custom recordings were used:

UCF50 is a widely adopted human action recognition dataset containing 6,618 video clips across 50 action categories,[27] sourced from real-world YouTube videos. These include everyday and high-movement activities such as running, basketball shooting, and handstand walking, providing valuable diversity for model training and generalization.

GMDCSA (General Multiview Dataset for Context-Aware Fall Detection) contains 1,752 annotated videos across 24 action classes,[28] of which focus on fall events and common daily actions in varied environments. With a dataset size of approximately 15 GB, it provides scenarios for validating the system's fall detection capabilities.

In addition to these datasets, SafeSight was evaluated using a set of recorded and live scenarios involving subjects performing both sedentary and non-sedentary actions. These real-time tests ensured that the system could operate effectively under practical deployment conditions, validating performance in privacy-sensitive environments and varying lighting or background conditions.

■ Result and Discussion

The SafeSight system was evaluated across multiple dimensions to assess its accuracy, responsiveness, and suitability for real-time monitoring of vulnerable populations. Results are presented across key modules, including scene context classification, action recognition, motion sensitivity, and system performance.

Scene Context Identification:

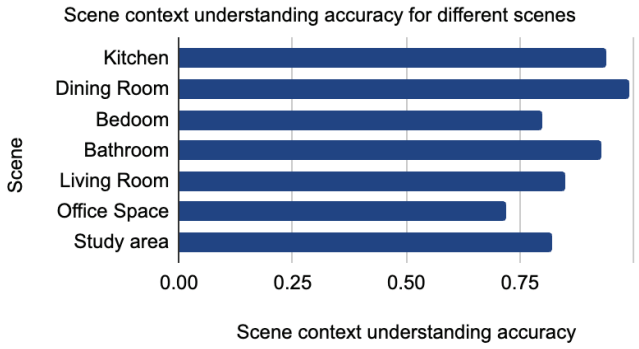

Figure 5: Scene context identification accuracy across various domestic environments, illustrating the highest accuracy in kitchens, dining rooms, and bathrooms. Performance slightly declines in office spaces and study areas, highlighting potential areas for model refinement. Overall, accuracy remains consistently high, supporting effective contextual classification in monitoring applications.

Scene context understanding accuracy varies across different environments. Figure 5 shows the accuracy for each context, with Dining Room and Kitchen yielding the highest performance (both >0.9), while Office Space and Study Area lag, likely due to visual ambiguity and overlapping furniture or layout. This indicates that the model performs best in visually distinctive domestic environments.

Action Recognizer:

Table 2 and Table 3 present detailed evaluation metrics for both sedentary and non-sedentary actions across live video scenes and benchmark datasets (UCF50, GMDCSA). The model achieved F1 scores of 0.867 for both action types in the various private spaces, indicating strong generalization in personal care contexts. Non-sedentary detection is more variable depending on the scenes, as the precision drops to 0.629, suggesting that similar body positions may confuse activity types in tight spaces.

Table 2: Performance metrics (accuracy, precision, recall, F1 score) for detecting sedentary and non-sedentary actions using live video datasets from various household scenes. Results show consistently high accuracy and precision in certain occupational areas, whereas some areas, like bathrooms, present challenges, highlighting opportunities for improvement in specific environments.

Dataset		Sedentary actions				Non sedentary actions			
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Live video across different scenes	Kitchen	0.765	0.771	0.867	0.815	0.700	0.729	0.762	0.744
	Bedroom	0.821	0.888	0.847	0.867	0.825	0.892	0.843	0.867
	Bathroom	0.794	0.867	0.832	0.849	0.693	0.629	0.600	0.614
	Living Room	0.719	0.763	0.800	0.781	0.707	0.814	0.688	0.745

Table 3: Evaluation of sedentary and non-sedentary action detection performance using GMDSCA and UCF50 datasets, respectively. Results demonstrate good accuracy, precision, recall, and F1 scores across both datasets, with slightly higher overall performance for non-sedentary actions.

GMDCSA Dataset Sedentary actions				UCF50 Dataset Non sedentary actions			
Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
0.712	0.768	0.732	0.749	0.793	0.729	0.800	0.814

The following bar plot in Figure 6 illustrates the average Euclidean distance between pose vectors representing a subject's posture across various actions, varying between static and dynamic states. The y-axis, "Average distance between two pose vectors," quantifies the degree of change from the individual's movement. Higher bars indicate a greater difference between the static and dynamic poses for a specific action, suggesting that there is significant movement or shifting. On the contrary, lower bars reflect actions with minimal pose variation when compared to a static pose. The x-axis labels each action, allowing for better comparisons and analysis of how different actions impact pose variation.

DOI: 10.36838/v8i1.98

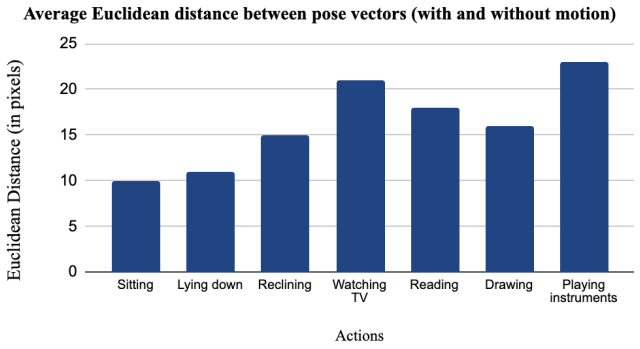

Figure 6: Average Euclidean distances between pose vectors across various domestic actions, indicating levels of movement from static to dynamic states. Higher distances, such as those in "Watching TV" and "Playing Instruments," reflect greater motion variability, whereas actions like "Sitting" and "Lying down" show minimal pose variations, illustrating lower movement intensity. This analysis aids in accurately identifying motion anomalies during sedentary behavior monitoring.

Motion Detection Sensitivity:

It measures the ability to detect movements accurately. It is calculated by computing the Euclidean distance between two pose vectors from the same time series but at different observation times. One of the sampling times is the current time, and the other sampling time is T seconds in the past (T is kept at 60 seconds and can be varied depending upon the scenario). The following Figure 7 shows how much the pose has changed over the sampling interval.

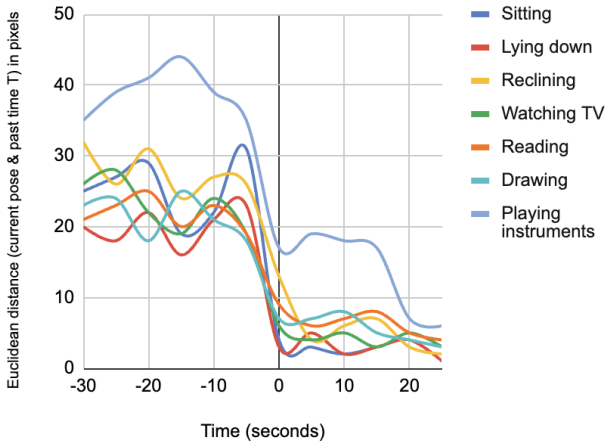

Figure 7: Motion detection sensitivity measured through Euclidean distance between current and past (60 seconds earlier) pose vectors over time. Distinct action patterns highlight varying motion intensities, with significant pose changes during actions like "Playing Instruments" and minimal movement during "Lying down," underscoring system accuracy in detecting and distinguishing subtle movements.

System Performance Metrics:

Latency, measured in seconds, refers to the time taken from the action to the alert. Depending on the scenario, the latency varies significantly. When an action happens within the context of the scene, such as cooking food on the kitchen stove, the system needs to compare Euclidean distances for a minimum of around thirty seconds before confirming that there is limited movement from the monitored individual. On the contrary, in cases where the actions are out of context, the action would be immediately flagged, triggering an alert within a few seconds. The latencies for these scenarios are directly compared in Table 4. In-context actions take longer to trigger alerts in the SafeSight system because they require extended monitoring to confirm anomalies, such as prolonged immobility, over a time window (32–180 seconds). Out-of-context actions, being inherently anomalous, trigger immediate alerts within 2 seconds.

Table 4: Latency measurements (in seconds) comparing system alert responses for out-of-context versus in-context actions. Immediate alerts (2 seconds) occur for anomalous, out-of-context actions, whereas contextually appropriate actions require extended monitoring (32–180 seconds) before triggering alerts.

	Cases when out-of-context actions happen	Cases when in-context actions happen
Latency (seconds)	2 seconds	32 to 180 seconds

Throughput is the number of frames or actions processed per second, which varies depending on how many models have to be run on the frame. Based on the particular scenario, throughput varies from 0.5 frames per second to 2 frames per second in the system. Additionally, the system's alert accuracy is dependent on the actions being tracked, differing between sedentary and non-sedentary actions as pictured in Table 5.

Table 5: System alert accuracy comparing sedentary versus non-sedentary actions. Non-sedentary actions achieve a higher accuracy (95%) compared to sedentary actions (87%), reflecting better reliability in dynamic scenarios.

	Sedentary actions	Non-sedentary actions
Alert Accuracy	87%	95%

Applications:

SafeSight has potential for broad applications beyond home-based monitoring for elderly individuals and children. In healthcare settings, it can assist in continuous patient monitoring, enabling early detection of fall events and tracking rehabilitation progress. In educational environments, SafeSight can enhance safety in classrooms and playgrounds, particularly for young or special-needs children. In industrial contexts, such as construction zones or chemical facilities, the system can detect hazardous activity patterns or accidents in real time. Additionally, SafeSight can be a powerful system in law enforcement and correctional institutions by providing monitoring to identify disturbances or violent events. Its privacy-preserving mechanisms make it suitable for public spaces, balancing the need for security with ethical surveillance.

Future Implementations:

Future versions of SafeSight can incorporate multi-modal sensing through depth cameras, LiDAR, and environmental sensors (e.g., temperature, gas, sound) to increase context awareness. Integrating these data streams with more advanced machine learning models can allow the system to work autonomously in complex scenarios, such as interactions between multiple people or within larger crowds. SafeSight could also be scaled to smart city infrastructures, where it would aid in public safety, disaster response, and population-level behavior monitoring in real-time. This evolution would support

deployment in transportation hubs, senior living communities, or post-disaster relief zones. Scaling SafeSight for smart cities and healthcare settings faces challenges, including high computational demands and network bandwidth constraints. Optimized algorithms and application-specific hardware will ensure efficient, large-scale deployment while preserving performance and privacy standards.

Despite visual masking, video monitoring may raise privacy concerns for users, especially in private settings. SafeSight will address these concerns by enhancing user control with customizable monitoring schedules and consent protocols. Although SafeSight's accuracy decreases in environments with visual ambiguity, future improvements will enhance scene differentiation and refine non-sedentary action classification in constrained spaces.

■ Conclusion

The SafeSight project presents a real-time, privacy-preserving monitoring system designed to safeguard vulnerable populations through intelligent scene understanding and behavioral analysis. By combining camera-based infrastructure with deep learning and vision-language models, SafeSight distinguishes between sedentary and non-sedentary actions, enabling rapid detection of potential emergencies with individual privacy protections. The architecture integrates context-aware modules for scene classification, object detection, motion analysis, and selective privacy masking. It effectively addresses key challenges in conventional systems that rely on wearable sensors. Experimental results show good performance, with F1 scores above 0.86 for sedentary action detection in live scenarios and high accuracy in context classification across common household settings. While current results validate the system's reliability, future development will focus on improving computational efficiency, enabling deployment at a larger scale, and incorporating multi-modal sensing for enhanced scene understanding.

■ Acknowledgments

The author sincerely thanks Murugan Sankaradas for the guidance and support throughout the research and implementations.

■ References

1. Rubenstein, L.Z. (2006). Falls in older people: Epidemiology, risk factors, and strategies for prevention. Age and Ageing, 35(suppl_2), ii37-ii41.
2. Saver, J.L. (2006). Time is brain—quantified. Stroke, 37(1), 263-266.
3. Patel, S., et al. (2012). A review of wearable sensors and systems with application in rehabilitation. Journal of NeuroEngineering and Rehabilitation, 9, 21.
4. Gjoreski, H., et al. (2017). How accurately can your wrist device recognize daily activities and detect falls? Sensors, 17(8), 1786.
5. Moore, Kevin et al. "Older Adults' Experiences With Using Wearable Devices: Qualitative Systematic Review and Meta-synthesis." JMIR mHealth and uHealth vol. 9,6 e23832. 3 Jun. 2021, doi:10.2196/23832
6. Chen, L., et al. (2013). Sensor-based activity recognition systems for ambient assisted living: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), 790-808.
7. Martín Abadi et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from https://www.tensorflow.org
8. Adam Paszke et al., PyTorch: An Imperative Style, High-Performance Deep Learning Library, 2019. https://arxiv.org/abs/1912.01703. Software available from https://pytorch.org/
9. Redmon, J., et al. (2016). You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640.
10. Bradski, G., The OpenCV Library, 2000, Dr. Dobb's Journal of Software Tools, Software available from https://opencv.org/
11. A. Bourke, J. O'Brien, and G. Lyons, "Evaluation of threshold-based tri-axial accelerometer algorithm for fall detection" Gait & Posture, vol. 26, no. 2, pp. 194-199, 2007.
12. K. Aminian and B. Najafi, "Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications," Computer Methods in Biomechanics and Biomedical Engineering, vol. 7, no. 4, pp. 263-273, 2004.
13. Tateno et al., Privacy-Preserved Fall Detection Method with Three-Dimensional Convolutional Neural Network Using Low-Resolution Infrared Array Sensor, 2020, Sensors, https://pmc.ncbi.nlm.nih.gov/articles/PMC7589648/
14. Gong et al., Enhancing Privacy with Optical Element Design for Fall Detection, 2023, Electronics Letters, https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/ell2.12995.
15. Meng et al., Privacy-Preserving Cameras for Fall Detection: Data Acquisition for Artificial Intelligence, 2024, Sensors, https://pubmed.ncbi.nlm.nih.gov/38657018/
16. Tateno et al., Hybrid SNN-based Privacy-Preserving Fall Detection using Neuromorphic Sensors, 2023, Proceedings of the Fourteenth Indian Conference on Computer Vision, https://dl.acm.org/doi/10.1145/3627631.3627650.
17. H. Zhao et al., "3D human pose estimation using single infrared image based on LiDAR data," Sensors, vol. 20, no. 11, 2020.
18. Camillo Lugaresi et al., MediaPipe: A Framework for Building Perception Pipelines, 2019, https://arxiv.org/abs/1906.08172
19. K. He et al., "Mask R-CNN," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397, 2020.
20. A. Radford et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021.
21. J. Lou et al., "Vision-language pre-training for multimodal understanding: A survey," ACM Computing Surveys, 2023.
22. R. Shokri et al., "Privacy-preserving deep learning," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.
23. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Language-Vision Alignment for Instruction Following with Multi-Modal LLMs. arXiv preprint arXiv:2304.08485.
24. "Moondream2: A lightweight VLM for on-device applications," GitHub, 2024. [Online]. Available: https://github.com/huggingface/moondream
25. Wang, J., Li, S., Zhang,., Zhang, L., & Han, Z. "YOLO-World: Real-Time Open-Vocabulary Object Detection." arXiv preprint arXiv:2304.00501, 2023.
26. Jocher, G. et al. "YOLOv8: State-of-the-art real-time object detection and pose estimation." Ultralytics, 2023. Available: https://github.com/ultralytics/ultralytics
27. UCF50 https://www.crcv.ucf.edu/data/UCF50.php
28. GMDCSA https://github.com/ekramalam/GMDCSA24-A-Dataset-for-Human-Fall-Detection-in-Videos

DOI: 10.36838/v8i1.98

Author

Mridula Murugan is a junior at South Brunswick High School with a passion for art, technology, and literature. She plans to major in electrical and computer engineering.