AI powered smart surveillance system using CCTV networks for crowd analysis, crime detection, and workforce monitoring

[1]Teja Govuri, [2]Ramindla Baba, [3]Pravallika Kothapalli, [4]G Lavanya, [5]D Sandhya Rani, [6]Mantesh

[1234]UG student, [5]Assistant Professor, [6]Associate Professor

[123456]Department of CSE

[123456]CMR Technical Campus Hyderabad, Telangana, India-501401

[1]tejabhoomreddy24@gmail.com, [2]babaramindla@gmail.com, [3]kothapallipravallika619@gmail.com, [4]lavanya.cse@cmrtc.ac.in, [5]davu.sandhya@gmail.com, [6]mantesh.cse@cmrtc.ac.in

***Abstract*—Because there are more cities and consequently more traffic on the roads because of the increased population density, accessibility to the look-out post is a challenge. Some places are currently using cameras. It can be a simple function of what continues – nothing to do with tricks. They are supposed to be seated and staring at the screen, hoping they can detect any threat that is there.**

Not very often. Mind_rambling, eye_rambling, reaction_time_slows down – Especially when it is expected to be quick. The person watching these screens is not always carly, nor "duly". Cameras sit in the same spot as before. Currently, they transmit their video streaming towards a program cap even the where it is detected if it is unusual. The act of being detected if it is unusual in its motion equals an alert straight towards the worker – without having to look at him fixedly. Fixedness of focus undergoes a change if there is a change that is conjectured. An era of nothingness happens, flowing smoothly, without a need for manual handling. Everything that the eye sees is from where "three pieces fit together."

In the motion presented to the petitioners, the number of individuals involved was estimated. In the areas where the spots for flags are located, people congregate. In the event of running or fighting, the alerts appear "immediately" - "Images saved for checking" later. Clocking IN and OUT will be Tape-recorded easily. Information Floods into systems, which forecast timetables regarding advance. It does everything on an incredibly simple screen!

“Nothing hidden, nothing fancy.”

Such an approach keeps people inside, taking the burden off their shoulders. Lots of work is left to computer video monitoring and alerting by sound, making decisions on what happens next is still in the hands of the staff. Paying attention is not an effort, and less is being demanded of it lately, while actions enable easy following when it matters most. Locations When “the world is packed with people” there is: CATCHING people disobeying rules is getting better, working every day with teams is also working with no flaw in it. Watching objects while pointing to them on frames. Strange patterns found through clever mathematical tricks. The cameras talk to each other without interruption. Lifting heavy loads: device vs. network. Change in space and time, motion. Unlocking its meaning. Certainly it could be that a crowd is full of people. It definitely helps with crowd control. In addition to that, the checking to see if they have the right safety gear "is something that happens as events are occurring"

I. Introduction

Virtual reality refers to the presence of computer-generated graphics. Essentially, it is a reality that technologically creates computer- simulated 3D images or environments. Also, it may have a similarity to the real world or be something completely other than that. Like this augmented reality, another option that is getting popularity these days. Consequently, individuals can expand their knowledge of the subject through free online courses. In addition, the hardware-software suite makes for an immersive environment for the user. Users also get a sense of presence or being in that environment with this immersion. A virtual reality device consists of multiple systems and hardware, such as gloves, a helmet, or goggles. Moreover, with the display of such devices, users are able to enter an interactive virtual 3D world. It is composed of five important components, namely immersion, interaction, sensory feedback, virtual world, and telepresence (i.e., presence in the environment). The first aspect is the generation of immersion and audio

II. RELATED WORK

Nowadays, taking photos isn't only about capturing scenes - smart tools help devices understand them too. Starting in 2020, Ishtiaq joined a small group developing models using DenseNet to spot tiny features. In bright settings, clear outcomes appeared without trouble, yet low light or foggy images created problems. That first push, however, led to progress; it influenced how machines later grasped motion and reacted on the fly. Not starting slowly, change arrived mid-2025 as Naidu's group turned on their gear - YOLOv8 paired with LSTM began spotting packed zones and strange motions fast. Yet performance dragged, held back by weak hardware.

From there, Dorothy teamed up with Priya, choosing simpler paths: compact Python scripts bridging TensorFlow and OpenCV for faster person detection. Movement tracking sharpened, running smooth, steady - until new, unknown sites enteredthe test loop. Sidewalks filled up once more during 2024 - Shwetha's group spent weeks tracing each camera spot street by street. Soon afterward, folks close to Bilade started doing the same things again. Warnings zipped across systems quickly, fed by CNN streams feeding into tools built on Django. Things ran without issue most of the time; now and then, just a small error crept.

Flickering visuals, then alerts popping up-trouble started piling fast. Out of nowhere, life appears in research reports. Not just inside lab walls do events unfold. Gears slip into daily patterns, soft, almost unnoticed. Moving fast matters.

III. PROPOSED FRAMEWORK

Another sudden shift emerges with the silent arrival of "aging cameras" gain new life through clever calculations that review the footage, no physical updates necessary. Rather, it employs what is already present, reducing costs without attracting attention. Footage appears at this point: insights are emerging over there, almost like following faint marks scattered across time. Days pass and some patterns emerge due to a guiding force that seems to by something called LSTM, noticing changes Others walk past. Only a warning is seen. When needed, silence otherwise. Speed increases even though the system remains small, running without being noticed. The images come apart, one by one, from these clips. Cleaning happens – Shrinking size, slicing HTML/Code, Static, Adjusting Brightness- for better view Now, smart systems analyze each picture, detecting.

1 : Faces shapes and motion as they move through. A shift occurs as soon as there is motion, and there destination equally against what gets left. Interaction creeps in gradually, then races ahead - depending on speed, how it is captured. "Stance emerges spontaneously, out of nowhere; space" stretches or closes between people. Piecing these moments across views builds a wider experiment of overlapping

perspectives: national regulatory Image Right from the beginning, the system applies LSTM to see how things change over time. Because when something happens, often the timing is the key, LSTM captures the rhythms present inside the data flow. Instead of freezing one frame, part of it watches groups move – how dense areas expand, where people move, second by second. Where there is a danger that the places will get too full, these changes will be useful "and keep them under constant observation. Over there, one could see the team track streaks of movement rather than individual particles. still images. In these cases, time continues to roll by with every gesture being incorporated into stories that portend danger. The idea of totaling hours is continued by comings: And goings can leave clues over a period of work. "The rate of arrivals, the flow of departures," theysketch the usual patterns without shouting. Just at the moment something strange takes place, the alarm goes off: clips head directly to the approved people. What the system does to the moments is what will ultimately determine if the moments will be used as input for the LSTM or if they will be rejected as a little off or as usual. Right at arrival, each piece of information displays a nicely packaged screen. Real-time monitoring without any delays regarding flow.
As a result, "make sense happens quickly" by design. The first thing you notice isn't loud – it is quiet to fill missing information in the blanks. From that moment on, patterns start to stick because the setup pays off. There are some patterns that are attention to what follows what. These pieces connect things together, not through magic but through tools ALREADY AVAILABLE. Most of the work is through eyes that were placed long before they had names. With everything moving so quickly, notifications just get received in time. Changes due to habits will begin to affect themselves, dangers are detected early. With delays slowing down, effort translates more smoothly. Where decisions involve real-time updates, they remain longer guesses every s()ingle time. Nothing stands out: because it simply works

IV. METHODOLOGY

The new way of building the security setup that uses intelligence starts here. It does not begin from the beginning. The AI-driven security setup uses camera systems and it pulls together advanced tools that can recognize patterns to study how people behave in groups, find actions that are not normal and maybe even track what is happening at the job site by looking at the footage from the camera systems. The AI-driven security setup does this one step at a time.The AI-driven security setup starts with cleaning up the raw video clips from the camera systems. Then it pulls out the details from the video clips that are worth watching. After that the AI-driven security setup uses neural networks to map the patterns over time. These neural networks are designed for sequences. They are called LSTM models. The LSTM models are a part of the AI-driven security setup that uses current camera systems. Warnings come up when something matches the things that the system is looking for which are set up deep inside the chain. The system has things that it checks for and when something fits these things that is when the warnings pop up. The chain is like a process and these triggers are deep inside it.

A. System Formulation

Let

$$V=\{V1,V2,V3,\ldots,VN\}$$

represent the set of CCTV video streams captured from different camera locations.Each video Vi is made up of lots of frames one, after the other.

$$Vi=\{fi1,fi2,fi3,\ldots,fiT\}$$

B. Video prep and feature extraction

Picture, by picture we make adjustments first. We change the size fade out the static and shift the balance. This helps to sharpen the results on.From there a special grid-like system is used to pick out shapes and movements. It figures out who is where and how things touch each other. A string of numbers is pulled from the data. This becomes what we call a feature vector. We use the feature vector to get the information we need from the pictures.

$$xt=f\theta(fit)$$

C. Temporal Modeling Using LSTM

What makes this setup work comes down to The system tracks changes over time. When it moves from one frame to the next the data goes through a part called an LSTM. This LSTM is made to hold on to information even when there are delays. It does not just look at changes. Instead it saves the parts of the information.It does this by making adjustments to the gates that control the information. At each step at a time the system updates its memory. It changes the way it remembers things at that time.

$$ht=LSTM(xt,ht-1)$$

At time t, the frame's past activity lives inside it. This hidden state carries earlier moments forward.This way of handling time lets time work, like this:

D. Module-wise Analysis

1) Crowd Analysis

When crowds start building up the system takes notice. If we watch what happens over a hours we can see how crowded a place gets. The predictions from the model are used to figure out how many people are there. If the number of people at the spot hits a threshold then the system sends out an alert because there are too many people packed into the spot. The crowds, at the spot are what the system is keeping an eye on.

$$Alert=\{1,ifDt>\delta 0,otherwise\backslash text\ \}$$

2) Crime Detection

When something moves in a way that's not normal it might mean something is wrong. The computer looks at what happens over time. Sorts it out. It does not say yes or no, it just says how likely something is to happen. It looks at what people do. Decides if it is normal or not. If something seems strange the computer will point it out. It will not make a big deal about it. The computer then decides if something is ordinary or odd. The machine looks at behavior. Decide what is ordinary and what is odd. It is always checking to see if something is ordinary or odd.

$$y\hat{}=Softmax(Wht+b)$$

3) WorkForce Monitoring

LSTM notes when people show up in video feeds. It logs the moment someone walks in, also captures when they leave. The system then figures out how long each person worked by using those times. They look at the times to see how long each person was working. This way they can tell how long each person worked using those times.

$$H=Texit-Tentry$$

E. Generate alerts and send notifications

Blinking lights appear the moment something out of the occurs. At that moment messages are sent out by email to the people who are allowed to see the blinking lights messages. The type of event that triggered the blinking lights is included in each blinking lights alert along with the time it happened. Pictures are also included with the blinking lights messages so someone can take a look, at what happened when the blinking lights event occurred.

F. Inference Phase

When we start figuring things out, new video clips go through the structure we learned from the CNN-LSTM.This process gives us a decision on what label to use for the video clips.

$$c^{\wedge}=\arg\max_{c} P(c \mid V)$$

V. EXPERIMENTS

Let us begin with the numbers. We need to look at how every single trial worked out over time. The system found clues when it saw things that it should have seen before. People, in charge kept an eye on the movements of the reading groups to make sure they were doing things correctly and they caught anyone who was not following the rules. At the time they also looked at what the staff were doing by watching the normal video feeds from the cameras that were put up all around the place.

A. Datasets

From a test we get footage coming in from street monitors and also from scenes that are set up in workplaces. We see what people do every day and also what happens when a lot of people are together, at events when we are looking at movement patterns. We look at the things that people normally do. We also pay attention to the things that are not normal which can show us when people are not following the rules. As workers go about their routines we get glimpses of what really goes on in offices and how they really work, which is interesting to see when studying movement patterns and movement patterns of workers.

From these video clips single pictures show up. They are labeled so we know what we need to keep an eye on. When we split these pictures the information starts to move. One part is used for practice and another part is set aside to check the answers later.

The progress we make comes from one part of the information and the conclusions we draw come from the part that is left over.

Teaching the computer does not happen in every part of the process it is the testing that happens in the corners, where the computer learns from the practice runs and the leftover information like the video clips and the single pictures.

B. Experimental Setup

The video goes into the main deep learning model. This model uses feeds from the monitoring systems. The video is cleaned up. Organized from the beginning. This way each piece of the video can be looked at closely. Then a scanner that uses vision looks at one picture at a time. As the video keeps going the important parts that are extracted from the video go into the motion path observer. The motion path observer looks at the video frame by frame. Spots any movement. It sees how things are moving and changing slowly over time. The video and the motion path observer work together to understand what is happening in the video.

A shape forms inside each part of the company. I watch meetings, I notice moments I note tasks. When I watch employees I often see that people focus on when they arrive rather than when they leave. The company is always. When we are in a learning phase we make adjustments after we make mistakes and we use a method called cross-entropy loss to learn from these mistakes and this is how the company learns from the cross-entropy loss.

C. Evaluation Metrics

When the training is finished the results come out. We use normal methods to evaluate how well the sorting jobs were done. First we look at how accurate the resultsre. Then we think about the precision of the results. After that we consider the recall. Finally we look at the F1- score. Each of these things shows us something about the results. What is important changes depending on which value is the most noticeable.

If the system marks things as correct this makes a big difference, in some of the metrics. This time we are paying attention to where the system made mistakes.

The last step takes all the attempts, puts them together and then makes a single shape out of them.

VI. RESULTS AND DISCUSSION

It shifted once scientists tested clever gadgets on old recordings. Following teams became simpler since the setup spotted strange behavior by tracking worker motions with almost no human input. Outcomes showed brief clips might begin understanding things alone after being worked through. Quiet scenes sparked warnings simply by tapping unseen thought levels below.

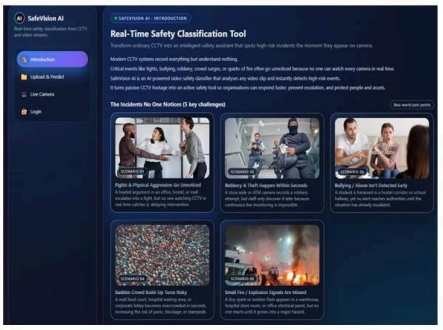

Figure 1: Real-Time Safety Classification Dashboard of the AI-Powered Smart Surveillance System

Right off, people show up clearly in live footage, moving through spaces. As they drift, their motion is tracked smoothly across zones. Then comes a steady scan of how packed each area turns out to be. When numbers climb too high, warnings pop up - fast. Overfilled corners or sudden clusters get flagged while still small. Instead of eyes glued to screens hour after hour, machines handle the watch. They never blink, never look away.

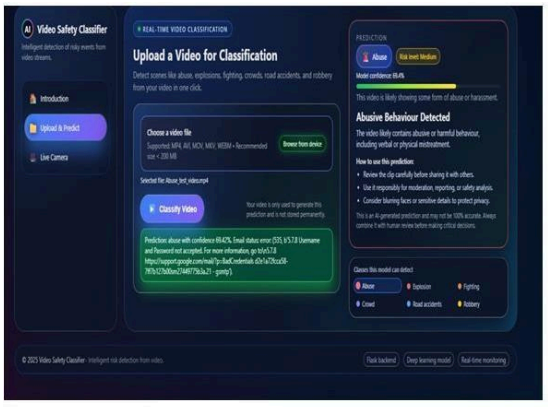

Figure 2: Video Upload and AI-Based Incident Classification Interface

Something jerky in the movement caught attention fast, simply because odd patterns got flagged right away. Right when weirdness appeared on screen, a message popped up – photo included. No hanging back for checks; warnings came through just as things happened, making reactions quicker. Dark corners or messy views sometimes confused it, yet performance climbed as lighting got better and code learned more. When images cleared up and detection tightened, advantages over old-style watching showed clearly. Footage rolls each morning when workers arrive through the front door. Motion sensors trigger recording devices automatically upon entry. Time stamps attach to every clip

showing who came in and when. Instead of signing sheets or swiping cards, faces show up on screen. Hours pile up frame by frame throughout the workday. Old methods fade as digital eyes take over routine checks. Silent cameras log shifts from start to finish every single day.

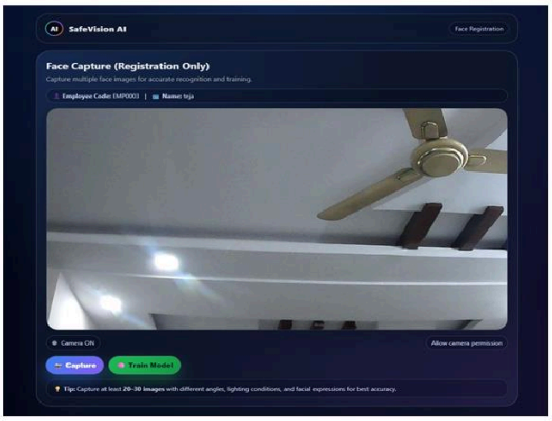

Figure 2: Video Upload and AI-Based Incident Classification Interface

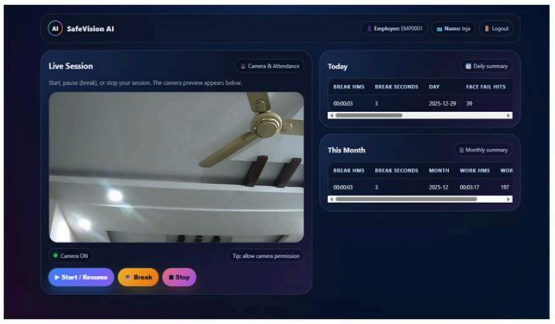

Figure 4: Live Camera Session Interface for Workforce Attendance Monitoring

When features work well, results still depend heavily on where cameras point, how light hits scenes, what details stand out clearly. Here is something not expected: mixing smart software into dated setups does better than people watching footage.

VII. CONCLUSION

Watch systems are changing the way things work now. Cameras are working with smart software to do a lot more than just record what is happening. They are actually studying what is going on in time instead of just sitting there doing nothing. When people walk by the cameras scan their movement patterns away. If someone does something it stands out really fast even when it is really busy. The main thing that makes this work is that we have control over what is going on using tools we already have. We get alerts early long before anything goes wrong.

Things start to run a lot once we get these warnings sooner and Watch systems are a big part of that.

When a crowd gets bigger something might happen. The system figures it out before things get out of hand. If people start acting, security finds out right away and they do not have to go through all the slow checks. You do not have to swipe cards or watch clocks because the cameras see you and log that you are there. This means you need machines.

Imagine if computers could learn what is normal and what is not. These systems can use the cameras that're already there so you do not need to buy new equipment. If something strange happens the system sends a warning before things get bad. The workers do not have to sit and watch the monitors all the time, they can just step in when they are needed to help with crowd issues like when a crowd grows. Speed here changes outcomes, especially when delays cost. Think sidewalks, workspaces, green areas - running this kind of watch beneath the surface. Faster steps ahead mean machines see better now. Somewhere up ahead, quiet streets might watch back.

References

[1] N. Shah, N. Bhagat, and M. Shah, ‘‘Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention,’’ Vis. Comput. Ind., Biomed., Art, vol. 4, no. 1, pp. 1–14,Apr. 2021.

[2] Autonomous Anomaly Detection System for Crime Monitoring and Alert Generation Jyoti Kukad, Swapnil Soner, Sagar Pandya

[3] V. Bilade, C. Fulaware, and A. Gaikwad, “Crowd Management Surveillance Using Artificial Intelligence and Deep Learning,” International Journal of Innovative Research in Modern and Practical Science (IJIRMPS)

[4] Redmon, J., & Farhadi, A. (2020). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.

[5] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934. [5] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable BagOf-Freebies Sets New State-Of The-Art for Real-Time Object Detectors. arXiv preprint arXiv:2207.02696.

[6] Girshick, R. (2024). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 1440-1448.

[7] Ren, S., He, K., Girshick, R., & Sun, J. (2023). Faster R CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, 28.

[8] Viola, P., & Jones, M. (2021). Rapid Object Detection using a Boosted Cascade of Simple Features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 511-518.

[9] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2022). Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)