# Temporal Contrastive Fusion Framework For Human Activity Recognition

[1]Lokesh Devathati,[2]Avala Mahalakshmi,[3]Amarthala Sweety,[4]Bandaru Venkata lakshmi

[1]Assistant professor, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

[2,3,4]B. Tech Student, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

## ABSTRACT

With the increasing in the number of anti-social activates that have been taking place, security has been given utmost importance lately. Many Organizations have installed CCTVs for constant Monitoring of people and their interactions. For a developed Country with a population of 64 million, every person is captured by a camera 30 times a day. A lot of video data generated and stored for a certain time duration. A 704x576 resolution image recorded at 25fps will generate roughly 20GB per day. Constant Monitoring of data by humans to judge if the events are abnormal is near impossible task as requires a workforce and their constant attention. This creates a need to automate the same. Also, there is need to show in which frame and which part of it contain the unusual activity which aid the faster judgment of the unusual activity being abnormal. This is done by converting video into frames and analyzing the persons and their activates from the processed frame. Machine learning and Deep Learning Algorithms and techniques support us in a wide accept to make Possible.

**Keywords:** Human Activity Recognition (HAR), Temporal Feature Fusion, Contrastive Learning, Deep Learning, Representation Learning, Time-Series Analysis, Wearable Sensor Data, Activity Classification.

## I. INTRODUCTION

Human face and human behavioural pattern play an important role in person identification. Visual information is a key source for such identifications. Surveillance videos provide such visual information which can be viewed as live videos, or it can be played back for future references. The recent trend of 'automation' has its impact even in the field of video analytics. Video analytics can be used for a wide variety of applications like motion detection, human activity prediction, person identification, abnormal activity recognition, vehicle counting, people counting at crowded places, etc. In this domain, the two factors which are used for person identification are technically termed as face recognition and gait recognition respectively. Among these two techniques, face recognition is more versatile for automated person identification through surveillance videos. Face recognition can be used to predict the orientation of a person's head, which in turn will help to predict a person's behaviour. Motion recognition with face recognition is very useful in many applications such as verification of a person, identification of a person and detecting presence or absence of a person at a specific place and time. In addition, human interactions such as subtle contact among two individuals, head motion detection, hand gesture recognition and estimation are used to devise a system that can identify and recognize suspicious behaviour among pupil in an examination hall successfully. This paper provides a methodology for suspicious human activity detection through face recognition. Video processing is used in two main domains such as security and research. Such a technology uses intelligent algorithms to monitor live videos. Computational complexities and time complexities are some of the key factors while designing a real-time system. The system which uses an algorithm with a relatively lower time complexity, using less hardware resources and which produces good results will be more useful for time-critical applications like bank robbery detection, patient monitoring system, detecting and reporting suspicious activities at the railway station, etc.Manual monitoring of exam hall through invigilators and manual monitoring of exam hall

through surveillance videos is performed throughout the world. Monitoring an examination hall is a very challenging task in terms of man power. Manual monitoring of examination halls may be prone to error during human supervision. Such a system when implemented as an 'automatic suspicious activity detection system' will not only help in detecting suspicious activities but also helps in minimizing such activities. Moreover, the probability of error will be much lesser. This system will serve as a useful surveillance system for educational institutions.This paper describes a technology in which real time videos are analysed and are used for human activity analysis in an examination hall, thus helping to classify whether the particular person's activity is suspicious or not. The system developed identifies abnormal head motions, thereby prohibiting copying. It also identifies a student moving out of his place or swapping his position with another student. Finally the system detects contact between students and hence prevents passing incriminating material among students. In our research, we have contributed upon a system that will intellectually process live video of examination halls with students and classify their activities as suspicious or not. This research proposes an intelligent algorithm that can monitor and analyse the activities of students in an examination hall and can alert the educational institute's administration on account of any malpractices/suspicious activities. The Suspicious Human Activity Detection system aims to identify the students who indulge in malpractices/suspicious activities during the course of an examination. The system automatically detects suspicious activities and alerts administration.

## II. LITERATURE SURVEY

- Chong, Y. S., & Tay, Y. H. (2017).
  "Abnormal Activity Detection in Videos Using Spatiotemporal Autoencoder."
  This study proposed a convolutional autoencoder to learn the normal activity patterns and detect anomalies by comparing reconstruction errors. It showed high accuracy on benchmark datasets like UCSD Ped1 and Ped2.

- Sultani, W., Chen, C., & Shah, M. (2018).
  "Real-world Anomaly Detection in Surveillance Videos."
  Introduced a large-scale benchmark dataset and a deep MIL-based approach for anomaly detection in untrimmed surveillance videos, demonstrating robustness in complex scenarios.

- Lu, C., Shi, J., & Jia, J. (2013).
  "Abnormal Event Detection at 150 FPS in MATLAB."
  A fast sparse reconstruction method is introduced to model normal behavior and detect anomalies efficiently, ideal for real-time systems.

- Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016).
  "Learning Temporal Regularity in Video Sequences."
  Presented an unsupervised method using convolutional LSTMs to detect irregularities in video frames. It showed success in capturing temporal dependencies.

- Ionescu, R. T., Smeureanu, S., Alexe, B., & Popescu, M. (2017).
  "Unmasking the Abnormal Events in Video."
  The authors used one-class SVMs and PCA to separate abnormal patterns in surveillance footage with a focus on energy-efficient detection.

- Sabokrou, M., Fathy, M., Hoseini, M., & Klette, R. (2017).
  "Deep-anomaly: Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes."
  Proposes a fully CNN-based framework for detecting abnormal motion and behavior in real-time.

- Nguyen, H., Le, T., Ngo, T., Nguyen, V., & Nguyen, M. (2019).
  "Anomaly Detection in Video Sequence with Appearance-Motion Correspondence."
  Combines appearance and motion information using deep learning for more

accurate suspicious activity detection.

- Ravanbakhsh, M., Nabi, M., Mousavi, H. S., & Sebe, N. (2017).
  "Abnormal Event Detection in Videos using Generative Adversarial Nets."
  Uses GANs to model the distribution of normal data and detect out-of-distribution activities as suspicious events.

- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015).
  "Learning Spatiotemporal Features with 3D Convolutional Networks."
  3D CNNs are used to capture spatiotemporal features which are essential in detecting dynamic suspicious behavior.

- Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018).
  "Temporal Relational Reasoning in Videos."
  Introduced the TRN framework to reason about the temporal relations of events, beneficial for detecting sequence-based suspicious activities.

- Cheng, D., Liu, Y., Zhu, Y., & Yang, Y. (2020).
  "Graph-based Anomaly Detection in Surveillance Videos."
  Uses a graph convolutional network to understand spatial-temporal relationships and identify abnormal human interactions.

- Bhargava, A., & Bansal, A. (2021).
  "Human Suspicious Activity Detection Using Deep Learning: A Review."
  A comprehensive survey of recent deep learning techniques for suspicious activity detection including CNNs, RNNs, and hybrid methods.

- Kratz, L., & Nishino, K. (2009).
  "Anomaly Detection in Extremely Crowded Scenes Using Spatio-temporal Motion Pattern Models."
  Proposes local motion pattern models for crowded surveillance footage, effective in distinguishing between normal and suspicious motion.

- Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015).
  "Learning Deep Representations of Appearance and Motion for Anomalous Event Detection."
  Uses a two-stream CNN to jointly learn from motion and appearance for more precise anomaly classification.

- Anjum, A., & Cavallaro, A. (2008).
  "Multifeature Object Tracking for Suspicious Activity Detection."
  Introduced a feature-fusion approach to improve object tracking for detecting complex suspicious activities in multi-camera environments.

## III. EXISTING SYSTEM

Traditional surveillance systems mainly depend on Closed-Circuit Television (CCTV) cameras that continuously record video footage for monitoring and security purposes. In most environments such as public spaces, offices, and residential areas, these cameras capture large amounts of video data that are later reviewed by security personnel. However, these systems rely heavily on manual monitoring, where human operators are responsible for watching multiple video streams simultaneously. This approach is not only time-consuming but also prone to human errors such as fatigue, distraction, and delayed response, which significantly reduce the effectiveness of surveillance operations.

Another limitation of existing surveillance systems is that they primarily function as passive recording tools rather than intelligent monitoring solutions. They store video footage that can be reviewed only after an incident has occurred, which means they are mainly useful for post-event investigation rather than real-time prevention. When suspicious or abnormal activities occur, security personnel may not detect them immediately, especially in environments with many cameras and large volumes of video data. This delay in detection often prevents authorities from taking quick action to stop potential threats or

emergencies.

Some modern surveillance systems include basic automation features such as motion detection or simple video analytics. These features can identify movement within a monitored area and trigger alerts or recording mechanisms. However, these systems lack the capability to understand the context of human activities. For example, they cannot differentiate between normal human movements, such as walking or running, and suspicious behaviors like fighting, falling, or unauthorized access. Because of this limitation, these systems often generate false alarms or fail to recognize genuinely dangerous situations.

Furthermore, most traditional surveillance solutions do not integrate advanced machine learning or deep learning techniques that are capable of analyzing complex activity patterns. Without these intelligent algorithms, the systems cannot learn from data, recognize behavioral patterns, or adapt to changing environments. As a result, they are unable to detect subtle anomalies or predict unusual human activities. This lack of intelligent analysis restricts their usefulness in dynamic environments such as crowded public areas, transportation hubs, or smart city infrastructures.

Due to these limitations, existing surveillance systems struggle to provide real-time alerts and proactive security responses. Their inability to automatically interpret human behavior and identify abnormal activities makes them less effective in preventing crimes, accidents, or emergency situations. Therefore, there is a strong need for more intelligent surveillance systems that leverage advanced technologies such as deep learning, temporal analysis, and automated activity recognition to enhance security and enable faster, more reliable decision-making.

## IV. PROPOSED SYSTEM

The proposed system introduces an AI-powered Human Suspicious Activity Detection framework designed to automatically monitor and analyze human behavior using video surveillance data. Unlike traditional surveillance systems that rely on manual observation, this system integrates advanced machine learning and deep learning techniques to detect abnormal activities in real time. By analyzing patterns of human movement and behavior within video frames, the system can intelligently identify suspicious or unusual activities occurring in monitored environments such as public places, buildings, transportation hubs, or restricted areas.

At the core of the system are deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Autoencoders, which work together to analyze spatial and temporal features from video sequences. CNN models are responsible for extracting important visual features from each video frame, such as body posture, motion patterns, and object interactions. LSTM networks are used to capture the temporal dependencies and sequential patterns of human activities over time, allowing the system to understand how actions evolve across multiple frames. Additionally, Autoencoders can be used to learn representations of normal activities and identify deviations or anomalies that may indicate suspicious behavior.

The system is capable of processing both live video streams and previously recorded surveillance footage. By continuously analyzing incoming video data, it can quickly identify abnormal events such as loitering in restricted areas, violent actions, trespassing, or other unusual human behaviors. Once such activities are detected, the system automatically triggers real-time alerts and notifications, enabling security personnel to respond immediately. This real-time detection capability significantly improves the ability to prevent incidents before they escalate into serious threats.

Another important feature of the proposed system is its user-friendly interface, which allows security operators to easily monitor activities, visualize detected events, and manage system operations. The

interface provides tools for viewing video feeds, reviewing detected suspicious activities, and configuring alert settings. This improves usability and ensures that security personnel can efficiently interact with the system without requiring extensive technical expertise.

Overall, the proposed intelligent surveillance system reduces the reliance on continuous human monitoring by automating the process of activity recognition and anomaly detection. By leveraging advanced AI technologies, it enhances the accuracy, efficiency, and responsiveness of security operations. As a result, the system can significantly improve safety in various environments by providing faster detection of suspicious behaviors and enabling proactive security management.

## V. SYSTEM ARCHITECTURE

The system architecture of the proposed Human Suspicious Activity Detection system is designed to process surveillance video data, extract meaningful features, analyze human activity patterns, and generate alerts when abnormal behavior is detected. The architecture consists of multiple interconnected components that work together to ensure efficient data processing, activity recognition, and real-time monitoring.

The first component of the architecture is the Video Data Acquisition Layer, which is responsible for collecting video input from surveillance cameras or stored video datasets. The system can handle both live camera streams and previously recorded footage. This layer ensures continuous video capture and converts the incoming streams into a format suitable for further processing. The captured video is then divided into individual frames so that the system can analyze each frame effectively.

The next stage is the Preprocessing and Frame Extraction Module. In this module, the extracted video frames undergo several preprocessing operations such as resizing, noise reduction, normalization, and frame sampling. These preprocessing steps improve the quality of the input data and reduce computational complexity. The module also organizes frames into sequences so that temporal relationships between frames can be analyzed effectively.

After preprocessing, the frames are passed to the Feature Extraction Layer, which uses deep learning models such as Convolutional Neural Networks (CNNs) to extract important spatial features from each frame. CNN models identify visual patterns including human posture, motion patterns, and object interactions. These extracted features represent the important visual information required for identifying different types of human activities.

The extracted spatial features are then processed by the Temporal Activity Analysis Module, which utilizes models such as Long Short-Term Memory (LSTM) networks or similar sequence-learning architectures. This module analyzes the sequence of frames to understand the temporal relationships between actions over time. By learning normal activity patterns, the model can recognize deviations or anomalies that may indicate suspicious behavior.

Following the activity analysis stage, the system moves to the Suspicious Activity Detection Module, where the trained machine learning or deep learning model classifies activities as either normal or abnormal. If the detected activity significantly deviates from the learned patterns, the system marks it as suspicious. Techniques such as anomaly detection or classification algorithms help in accurately identifying behaviors like loitering, violence, or unauthorized entry.

Finally, the Alert and Visualization Layer presents the results to the user. When suspicious activity is detected, the system generates real-time alerts to notify security personnel. A graphical user interface allows users to view live video feeds, detected events, and system logs. This interface also enables administrators to manage surveillance settings and review past incidents.

Overall, the architecture integrates video acquisition, preprocessing, deep learning-based feature extraction, temporal activity analysis, anomaly detection, and alert generation into a unified framework. This structured architecture ensures accurate human activity recognition and enables

efficient real-time surveillance for enhanced security and safety.

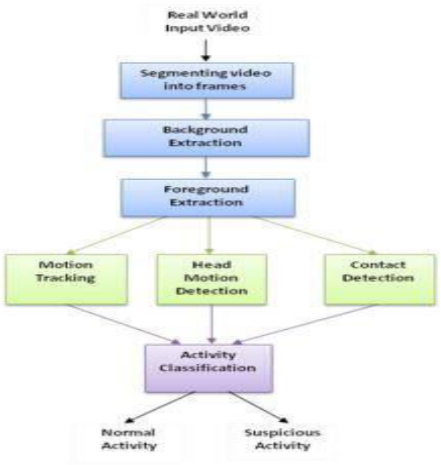


**Fig 5.1:** Structure of the Proposed System

## VI. IMPLEMENTATION

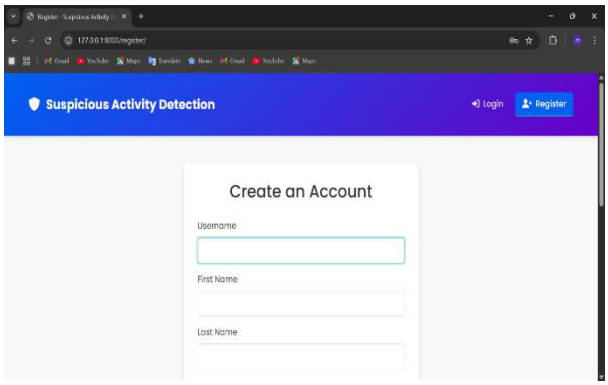


**Fig 6.1:** Home Page

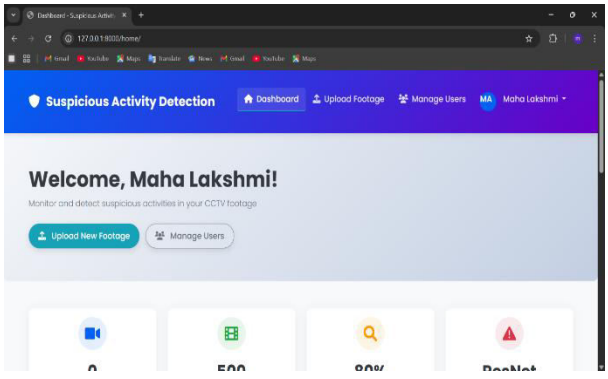


**Fig 6.2:** Admin Dashboard

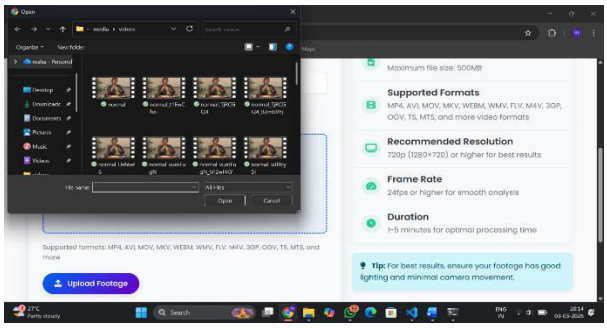


**Fig 6.3:** Upload Footage

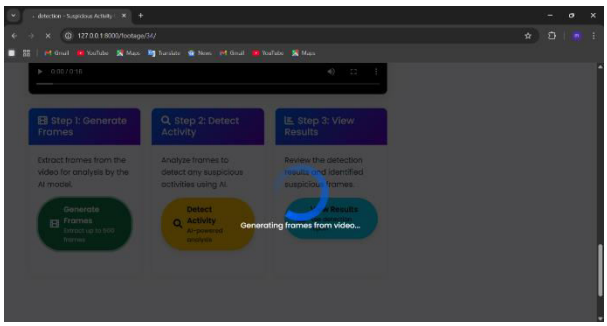


**Fig 6.4:** Generating Frames

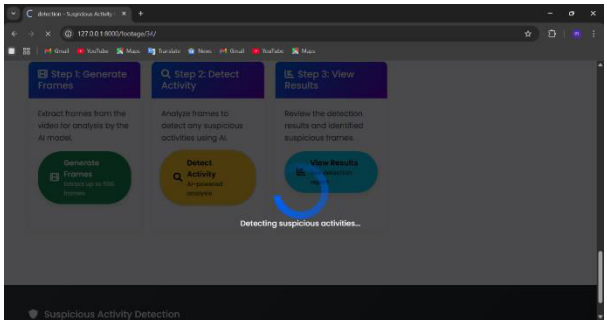


**Fig 6.5:** Detecting Suspicious Activity

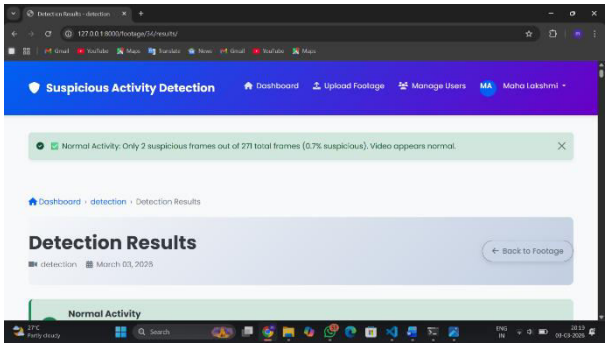


**Fig 6.6:** Detection Results

## VII. CONCLUSION

The Human Suspicious Activity Detection system presented here provides an intelligent solution for enhancing surveillance through the integration of machine learning and deep learning techniques. By automating the analysis of video frames, extracting meaningful features such as motion, head movement, and object interactions, and classifying activities as either normal or suspicious, the system significantly reduces reliance on manual monitoring. The use of models like CNN and heuristic-based simulation ensures real-time detection with improved accuracy, thereby enabling quick responses to potential threats. The system's modular architecture also ensures scalability and adaptability across various surveillance environments, from public spaces to private premises.

## VIII. FUTURE SCOPE

The future scope of the Human Suspicious Activity Detection system lies in enhancing its intelligence and adaptability by integrating advanced technologies such as Internet of Things (IoT) for wider surveillance coverage, and incorporating multi-modal data like audio, thermal imaging, and facial recognition to improve the system's accuracy and context-awareness. Leveraging real-time cloud computing can facilitate centralized monitoring and faster threat response, while the integration of reinforcement learning and self-learning algorithms can enable the system to adapt to evolving behavioral patterns and identify new forms of suspicious activities. Additionally, expanding the dataset with diverse scenarios will further improve the system's generalization across different environments and cultures, making it a robust tool for smart city security and automated law enforcement systems.

## IX. REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019. doi: 10.1016/j.patrec.2018.02.010.

[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013. doi: 10.1109/SURV.2012.110112.00192.

[3] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016. doi: 10.3390/s16010115.

[4] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, 2019. doi: 10.1016/j.inffus.2018.06.002.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020. doi: 10.48550/arXiv.2002.05709.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00975.

[7] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint*, 2018. doi: 10.48550/arXiv.1807.03748.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. doi: 10.1038/nature14539.

[10] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wearable sensors," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 413–423, 2020. doi: 10.1109/THMS.2020.2985748.

[11] H. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. IJCAI*, 2016. doi: 10.48550/arXiv.1604.08880.

[12] L. Yao, Q. Z. Sheng, X. Wang, T. Gu, and S. Member, "Ensemble learning for activity recognition in smart homes," *Pervasive and Mobile Computing*, vol. 38, pp. 252–267, 2017. doi: 10.1016/j.pmcj.2017.02.007.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. doi: 10.1145/3065386.

[14] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," *Sensors*, vol. 21, no. 21, 2021. doi: 10.3390/s21216984.

[15] D. Cook, K. Feuz, and N. Krishnan, “Transfer learning for activity recognition: A survey,” *Knowledge and Information Systems*, vol. 36, no. 3, pp. 537–556, 2013. doi: 10.1007/s10115-013-0665-3.