Deep Learning-Based Control System for Context-Aware Surveillance Using Skeleton Sequences from IP and Drone Camera video

Vasavi Sanikommu
Computer Science and Engineering
Velagapudi Ramakrishna Siddhartha
School of Engineering
Siddhartha Academy of Higher
Education, Vijayawada, India
vasavi_movva@vrsiddhartha.ac.in

Sobhana Mummaneni
Computer Science and Engineering
Velagapudi Ramakrishna Siddhartha
School of Engineering
Siddhartha Academy of Higher
Education, Vijayawada, India
sobhana@vrsiddhartha.ac.in

Novaline Jacob
Advanced Data Processing Research
Institute, Department of Space, Indian
Space Research Organization (ISRO)
Hyderabad, India
novalinejacob@adrin.res.in

Emmanuel Sanjay Raj K.C
Advanced Data Processing Research
Institute, Department of Space, Indian
Space Research Organization (ISRO)
Hyderabad, India
sanjay@adrin.res.in

Bhartendra Kumar
Computer Science and Engineering
Velagapudi Ramakrishna Siddhartha
School of Engineering
Siddhartha Academy of Higher
Education, Vijayawada, India
jrf1@vrsiddhartha.ac.in

Radha Devi Pullur Variam
Advanced Data Processing Research
Institute, Department of Space, Indian
Space Research Organization (ISRO),
Hyderabad, India
director@adrin.res.in

Abstract: *Human Activity Recognition (HAR) combined with face recognition is set to play a decisive role in next-generation surveillance systems. This work presents a hybrid methodology that integrates deep learning and machine learning models for recognizing multi person activities and faces. The work is structured into two different parts: face recognition and human activity recognition. For face recognition, faces are detected using the state-of-the-art Multi-Task Cascaded Convolutional Neural Network (MTCNN) model, followed by key point extraction with the FaceNet model. The extracted embeddings are classified using a Support Vector Machine (SVM) to identify individuals. SVM model achieved classification accuracy of 0.99. For activity recognition, an ensemble model is employed to classify six activities: walking, standing, sitting, punching, kicking, and crawling. The YOLOv8 large pose model is used to extract human skeletons, which are then fed into the ensemble machine learning model for classification. This integrated system demonstrates promising performance for real-time surveillance applications that detect and recognize the multi person activity and track the person. Generation of summary report is one of the most important phase of this work where the location details of a person is stored along with activity being performed by the person. If abnormal activity is recorded, then the system will generate the early warning system that helps for better surveillance purposes.*

Keywords: *Computer vision, human activity recognition, face recognition, crowd video surveillance, deep learning, feature extraction.*

Received February 4, 2025; accepted June 23, 2025
https://doi.org/10.34028/iajit/22/5/6

1. Introduction

Human Activity Recognition (HAR) task is very challenging and prominent task in the field of the computer vision [36]. In the recent years HAR attracted and gained the attention of the researchers in this area, recognizing human actions holds significant value, as it provides insights into how individuals interact with their environment and convey emotions [4]. HAR play important role in Surveillance, industrial-automation, robotics, ecommerce, sign language recognition [42]. Human action recognition in low resolution and illuminations conditions is even challenging task [35]. Human activity recognition on 3D requires transformers that will estimate correlations among human body joints is explained by Wang *et al*. [38]. Recognizing hand gesture is crucial in human computer interaction. Kinect sensor can be used to recognize gestures in uncontrolled environments [27]. Online clothing stores require image capture and dependencies identification of a person for a virtual try on the dress fittings [26]. Collecting the datasets for training and testing phase is equally important in human activity recognition because exiting methods depend on detecting usual patterns [21] and with spatio-temporal sequences [20]. As such light weight models with less computation are introduced for human activity recognition [43]. Machine learning algorithms depend on hand-crafted feature extraction process that learns from the given dataset and also to make predictions on the output. Deep learning algorithms performs many iterations to learn both low level and high level features from the given dataset and to make final predictions. In recent years, substantial progress has been achieved in skeleton-based action recognition, outpacing traditional hand-crafted approaches. This advancement is largely attributed to the remarkable developments in deep learning techniques [2]. Integrating face recognition with activity recognition further enhances the effectiveness of surveillance systems, from this task can recognize the person and activity simultaneously.

In daily life human perform many activities but every activity is related to the core activities such as standing, sitting, running. Such activity can determine the anomaly that can help to the abnormal activity detection, human activity anomaly detection is one the important and challenging task for the surveillance, robotics, and other human machine interaction task. This task has challenges during variations in clothing, changes in lighting and background noise. These factors significantly impact the HAR work and model performance. To address these issues, both visual and non-visual sensors can be employed. In the visual sensor domain, options include infrared cameras, thermal cameras, multispectral and Red, Green, Blue-Depth (RGB-D) cameras, while in the non-visual domain, audio signals, accelerometers and electro-thermal sensors are used. Each sensor type offers unique advantages and limitations, making sensor selection critical for effective implementation [8]. Traditionally, action recognition has utilized different types of Two-Dimensional Convolutional Neural Network (2D-CNN) and Three-Dimensional Convolutional Neural Network (3D-CNN) [19]. A generic Convolutional Neural Network (CNN) that exploits Red, Green, Blue (RGB) image sequences is required for video understanding [17] and for action recognition [34]. Frame rate at which the video is processed for human activity recognition plays essential role in capturing spatial semantics [9]. Few classifiers extracts features in specific to the query in hand and accordingly will recognize the action in the given video stream [32]. Lie theory captures the similarities within streams and can be used for pose estimation [40]. Apart from skeleton, layouts that gives depth sequences of human body can be considered for patient activity recognition [3]. Fusion of RGB and optical flow can reduce the usage of trainable parameters in neural networks and also with improved performance [10] and computational cost [37]. These approaches have achieved remarkable outcomes using diverse techniques. Skeleton data based human activity recognition models provides a precise topology-based human body model via bones and joints and are often more demanding in computation and less dependable in handling complex backgrounds and changing conditions. Challenges such as variations in body size, viewing angles, and motion speeds further add to their limitations.

Face recognition is a specific task within visual pattern recognition. While humans effortlessly interpret visual patterns through their eyes and process them into meaningful concepts in the brain, computers view images and videos as matrices of pixels. The challenge for machines lies in identifying what a particular section of this data represents [16]. Face recognition activity will determine the identity of the person in the detected face, making it a specialized classification problem. The human face is a complex structure that conveys important information, such as expressions, emotions, and distinctive features. Accurately analyzing these facial details for the purpose of attendance monitoring is challenging and often requires significant time and effort. In recent times, several facial recognition algorithms have been developed and successfully applied in automatic facial recognition system and implemented for the attendance management systems [31] and displaying on a Liquid Crystal Display (LCD)/Light Emitting Diode (LED) screen for attendance monitoring of a student [33]. Heterogeneous face data is generated because of sketch artists, spectral differences during face capture [18], intra class variation during face capture by multiple cameras [5]. Deep generative algorithms continue to emerge to handle such heterogeneous face data [11], and cross domain matching data [12], while existing ones are being refined or combined with other techniques to enhance the performance of facial recognition systems [14].

1.1. Motivation

This research is driven by the need for practical applications that integrate human activity recognition with face recognition using artificial intelligence, a field that remains relatively underexplored in real-world scenarios.

1.2. Objectives

- To develop deep learning system for multi person activity and face recognition system in real time for secured surveillance system.
- To create face dataset with Internet Protocol (IP) camera, Closed-Circuit Television camera (CCTV) and Drone based.
- To test the performance of the ensemble model with MTCNN and YOLOV8 for activity and face recognition.

1.3. Contributions

- Created a dataset for face and activity recognition tasks using drone, CCTV footage, Infrared (IR) and IP cameras.
- To use this dataset for training, testing and to validate the model, enabling a more accurate evaluation of its performance.
- Proposed ensemble based deep learning model to detect person, recognize the person and their activity in real time environment.
- Developed context aware sub system to generate alerts during suspicious activity

1.4. Research Gaps

- Existing research has primarily focused on face recognition from limited environments. As such there is a gap in understanding how face recognition and activity recognition algorithms perform on IP camera and drone camera images.
- While there is considerable research on face recognition using skeleton data there is lack of precise model that is suitable for robust environments with complex backgrounds and changing conditions.
- There is a gap in research that systematically investigates on context aware alert generation systems for multiple capturing systems.

1.5. Organization

Section 2 is all about previous work related to deep learning and context aware systems such as activity recognition and face recognition. Section 3 describes the methodology of the proposed work and section 4 provides the results, discussions and finally conclusions and later works.

2. Literature Survey

This section presents a concise overview of the existing works on activity recognition and face recognition techniques and their importance. Advanced human activity recognition systems primarily depend on raw RGB data for anomaly detection [6]. Recently, Sultani *et al*. [30] explained supervised abnormality detection system using RGB data, which assesses abnormality by estimating scores for each video frame, focusing only on frame-level anomalies without considering pixel or target-level details. In semi-supervised learning, autoencoder-based methods play a key role. Hasan *et al*. [13] used CNN-based autoencoders to capture temporal patterns in Histogram of Oriented Gradients-Histogram of Optical Flow (HOG-HOF) features. These autoencoders are trained on normal data, so test clips that match this data are reconstructed with low error, while unusual data produces higher reconstruction errors, making anomaly detection possible through error analysis. Chong and Tay [7] proposed a similar method that combines CNN and Long Short-Term Memory (LSTM) for encoding and decoding both spatial and temporal data. However, this approach doesn't clarify the nature of the anomalies it detects. Other methods [1, 44] used background subtraction as a pre-processing step on raw data to focus on moving objects. Yet, background subtraction struggles in scenes with cluttered backgrounds or when using pan-tilt-zoom cameras. In contrast, human pose detection remains effective in handling these challenges.

Yadav *et al*. [41, 42] reported Channel State Information Time-series model (CSITime) architecture and validated the model using Stanford-like WiFi (StanWiFi) and Activity Recognition with Information from Links (ARIL) datasets. Salehzadeh *et al*. [29] explained a deep learning model to identify Electroencephalography (EEG) artifacts associated with physiological. Their method utilized the combined capabilities of different CNN to achieve effective classification of human activities. Detecting actions such as jaw clenching and head or eye movements, which are often challenging with sensory technologies commonly used in activity recognition, is handled more efficiently by this model. Liu *et al*. [22] developed an approach for recognizing human activities using smartphone sensor data. Features were extracted from the data, and machine learning classification models were applied to analyze the activities. The performance of the method was ultimately evaluated using a CNN.

Table 1. Summary of the related work.

Existing work	**Methodology**	**Advantages**	**Disadvantages**
[6]	Autoencoder Architecture that uses RGB data for anomaly detection	The use of autoencoders produced better prediction of Anamoly	Usage of k-means clustering dpends mailny on the intial values of cluster centers. also the algorithm is tested on publicly available datasets
[30]	Supervised abnormality detection system using RGB data	Frame-level anomalies are tested in this system	Their system will not consider pixel or target-level details
[13]	CNN-based autoencoders to capture temporal patterns in HOG-HOF features.	Low error	Not tolerant with untrained data
[7]	Ensemble of CNN and LSTM	Works for spatial and temporal data.	Does not clarify the nature of the anomalies it detects
[41, 42]	Inception Time network architecture	Model is evaluated with low quality datasets	AUC measure alone cannot coclude on the model accuracy
[29]	Multivariate Gaussian model	Good accuracy in classification of human activities using optical flow algorithms.	Limted datasets such as UCSD-PED2 is used to test the model accuracy
[22]	Deep electro encephalography learning method	Deep learning framework to evaluate classification of EEG artifacts	Mental state of the user alone cannot predict activity recognition
[15]	Compared machine learning algorithms for facial recognition.	Their approach is integrated into mobile applications for improved accuracy and performance	Their work is limited to walking and running activites
[39]	Face recognition by integrating SVM with Particle Swarm Optimization.	This system handles non-linear problems and PSO's efficiency in optimizing SVM parameters	Their system is dependent on machine learning algorithms and may have overfitting problems when dataset is changed

Kremic and Subasi [15] compared different machine learning algorithms for facial recognition. Their approach involved reading images, skin color detection, converting RGB to grayscale, histogram processing, and classification. They have integrated these methods into mobile applications for improved accuracy and performance.

Wei *et al*. [39] explained face recognition by integrating SVM with Particle Swarm Optimization (PSO). This combination leverages SVM's capability to handle non-linear problems and PSO's efficiency in optimizing SVM parameters. The authors applied this method to the Facial Recognition Technology (FERET) human face database and compared its performance with traditional SVM and Backpropagation Neural Networks (BPNN). The outcomes demonstrated that the PSO-SVM method attain good accuracy indicating its effectiveness in face recognition tasks [39]. Table 1 summarizes the existing works and reports advantages and disadvantages of each method.

3. Methodology

3.1. Proposed System

Figure 1 illustrates the overall flow diagram of the methodology, providing an overview of the workflow for this study. Two separate datasets were collected for two distinct tasks: Face recognition and activity recognition. For face recognition, a custom dataset was created by collecting images from 27 individuals, with a minimum of 200 images per person. The face recognition model was trained using this dataset. Simultaneously, a dataset for activity recognition was prepared by collecting data for six different activities. For each activity class, approximately 200 images were extracted from video frames. This workflow highlights the systematic approach taken to collect, pre-process, and utilize the datasets for training the respective models.

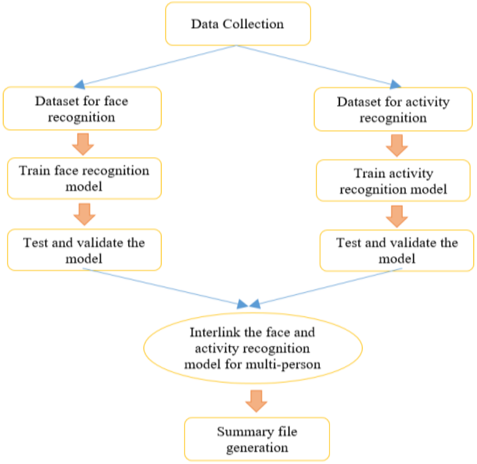

Data Collection
Dataset for face recognition
Dataset for activity recognition
Train face recognition model
Train activity recognition model
Test and validate the model
Test and validate the model
Interlink the face and activity recognition model for multi-person
Summary file generation

Figure 1. Block diagram for the face and activity recognition model.

3.2. Activity Recognition

Human activities can be broadly classified as if they are performed by a single person or a set of persons. Walking, crawling, sitting, standing are single person activity, whereas kicking and punching involves more than one person. Deep learning algorithms help in extracting these relationships between a single or multiple persons.

3.2.1. Dataset for Activity Recognition

For this HAR study, the model is trained to recognize six classes: walking, sitting, crawling, standing, punching, and kicking [23, 25] as shown in Table 1. Each class requires a dataset where only one individual appears in each frame, as the presence of multiple people can interfere with the classification process. As shown in Table 2, 27 people (both faculty and students of our institute) consisting of 200 images for four activities are collected using drone. To ensure effectiveness, the dataset was captured under diverse conditions. Images include variations in facial angles and expressions to account for different real-world scenarios. This comprehensive dataset is designed to effectively train the model to recognize faces accurately, even under different challenging conditions. Database containing three activities (walking, sitting, standing) for ten people are collected from college surveillance camera stored in digital video and network video recorders. The video is recorded using a static camera placed at different corridors with 30 frames per second. Histogram based key frame extraction is performed to collect key frames. Further these frames are scaled to 256X256 resolution. FLIR IR camera is used to collect video footage during night times in the department corridor for two activities walking and crawling. Nearly 25 images with 80X60 resolution is considered for testing the proposed model.

Table 2. Dataset description.

S.No	Source	Number of images	Class
1	Rajput *et al*. [25]	150	Standing, Sitting, Walking
2	Naik and Naik [23]	150	Kicking, Punching
3	Drone data	150	Standing, Sitting, Walking, Crawling
4	CCTV footage and IP caemra	200	Standing, Sitting, Walking,
5	IR camera	25	Walking, Crawling

Following data collection, pre-processing is equally critical. During this stage, images containing multiple individuals or unrelated activities are removed to ensure accurate classification. Background subtraction is also included in this phase to recognize the moving activities walking, crawling. To ensure data consistency, frames that did not maintain homogeneous classes are removed.

3.2.2. Action Recognition Using Ensemble Model

HAR is a complex task that requires sophisticated techniques to achieve accurate results. This research work employed a fusion deep learning and machine learning model to recognize various human activities. The first step involved collecting and preprocessing the dataset, where the videos are meticulously filtered out that did not maintain homogeneous classes to ensure data consistency.

Subsequently, YOLOv8 pose estimation model is introduced, which was instrumental in extracting human skeletal structures and tracking individual movements within the video sequences. The YOLOv8 model is available in various versions, such as large, medium, small, and Nano, each offering unique advantages and trade-offs. For instance, while the YOLOv8 large model excels in detecting and extracting skeletal features with high accuracy, it operates at a slower processing speed, making it less efficient compared to the YOLOv8 Nano model, which, although faster, sacrifices some degree of precision. For this study, YOLOv8 nano (YOLOv8n) model is used, as its superior accuracy in skeletal detection was deemed crucial for the reliability of the activity recognition system, despite the longer processing times. Figure 2 illustrates the extracted 17 key points using YOLO model, for training the ensemble model.

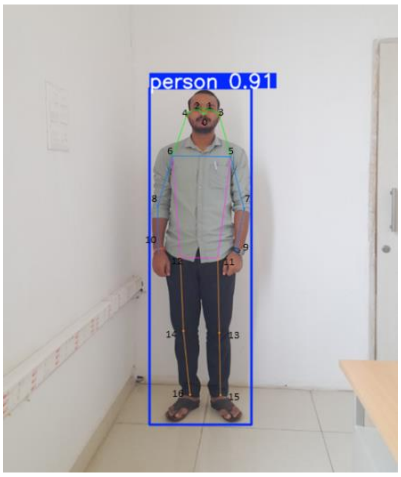


Figure 2. YOLOv8 model pose detection with 17 keypoints.

The skeleton data was extracted, and all key points were saved in .csv file format for subsequent use in training the pose recognition model. To classify the poses, four different classifiers: random forest, SVM, naive bayes and K-Nearest Neighbors (K-NN) are used. Following are the reasons for choosing these classifiers for human activity recognition. Random forest algorithm is robust to noise, class imbalance (as can be seen in Table 2) and can handle over-fitting of the features because of ensemble of decision trees being employed. SVM algorithm is also suitable for classifying class imbalance data. Usage of non linear kernels can distinguish between different activities especially kicking and punching activities. Forming clusters for six activities using K-NN is simple and suits well for medium datasets. Even though there is little bit coherence for kicking and punching activities, naïve bayes is well suited for resource constrained environments and as such used in the current research work. Four models are used for activity recognition to ensure best method for HAR.

After classification by each of these classifiers, a soft voting mechanism was implemented, wherein the final pose classification was determined by calculating the mean of the votes, thereby assigning a class to the corresponding activity. Figure 3 showing the methodology for the activity recognition using ensemble method.

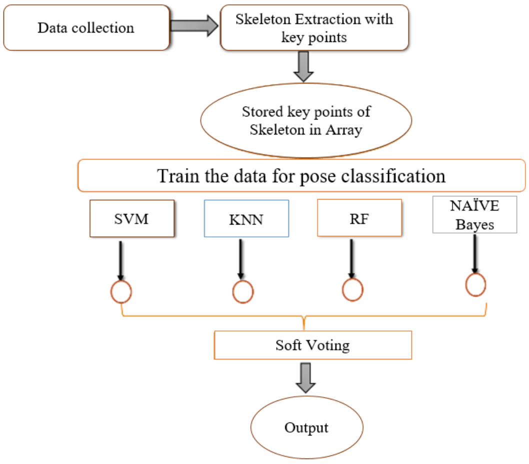

Data collection
Skeleton Extraction with key points
Stored key points of Skeleton in Array
Train the data for pose classification
SVM
KNN
RF
NAÏVE Bayes
Soft Voting
Output

Figure 3. Methodology for activity recognition.

3.2.3. Face Recognition

For face recognition task, custom dataset of 27 individual people and for each class around 200 images are collected, to ensure effectiveness of the model. This dataset was captured under diverse conditions. Images include variations in facial angles and expressions to account for different real-world scenarios. This comprehensive dataset is designed to effectively train the model to recognize faces accurately, even under different challenging conditions. For HAR, the model is trained to recognize six classes: Walking, sitting, crawling, standing, punching, and kicking as shown in Table 2. Each class requires a dataset where only one individual appears in each frame, as the presence of multiple people can interfere with the classification process. Following data collection, pre-processing is equally critical. During this stage, images containing multiple individuals or unrelated activities are removed to ensure accurate classification.

3.2.4. Face Detection

Face detection was carried out utilizing the pre-trained MTCNN model as shown in Figure 4, which is well suited for face detection tasks. The model exhibited a high degree of accuracy in identifying faces under varying angular tolerances our focus is to detect the faces more than 15-degree tolerance, effectively handling faces captured from different orientations. Additionally, the performance of the MTCNN model was tested across a range of distances, including short-range, mid-range, and long-range scenarios using dataset collected from drone and different cameras. This evaluation confirmed the model's capability to maintain reliable face detection across diverse spatial scales, demonstrating its effectiveness for applications involving variable distances and viewpoints.

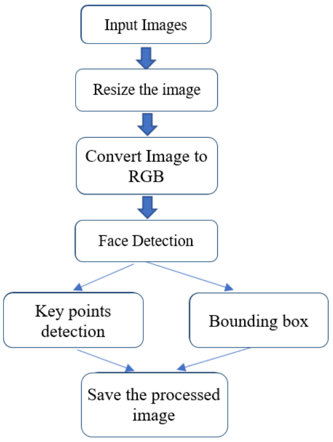


Figure 4. Face detection flowchart.

3.2.5. Face Recognition Workflow

Facial recognition has become a leading biometric method for identity verification and is widely used in fields like military, finance, public security, and daily applications. This work employs a hybrid model for both detection and recognition of faces from different situation, here utilizing MTCNN for face detection and FaceNet pre-trained model for embedding generation as shown in Figure 5. First, MTCNN detects faces within the dataset, identifying bounding boxes and key facial landmarks to crop and resize each face image to 160x160 pixels, meeting the input requirements of the embedding model. Next, FaceNet model extracts a 512-dimensional feature vector, from each face, capturing unique characteristics essential for classification. These embeddings, along with their associated labels, are split into 80/20 training and testing set. Classification is achieved with a SVM classifier using a linear kernel, configured to generate probability estimates, which adds robustness to the recognition process. The classifier model is trained based on the embeddings extracted from the FaceNet model and then evaluated for accuracy assessment using testing set. This approach integrates deep learning for feature extraction with machine learning for classification, establishing a robust system for precise and dependable face recognition.

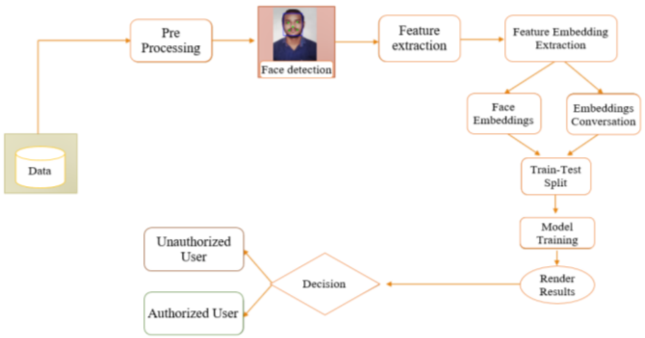


Figure 5. Face recognition flow diagram.

3.3. Evaluation Metrics

The performance of the ensemble model for activity and face recognition model was evaluated using accuracy, recall, precision, F1-score as shown in Equations (1), (2), and (3). To understand the class-wise performance, precision, recall, and F1-score were utilized. Additionally, a confusion matrix was plotted to provide detailed insights into true positive and false positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where *FP* means false positive, *FN* means false negative, and means true positive.

Cross validation score with five-fold is calculated for the iterative training and validating the model using Equations (4) and (5).

$$Fold\ Accuracy_i = \frac{Cross\ Prediction_i}{Total\ Instances_i} \qquad (4)$$

$$Mean\ CV\ Accuracy = \frac{1}{k}\sum Fold\ Accuracy_i \qquad (5)$$

3.3.1. Activity Recognition Performance

Figure 6 presents the model's performance across six activity classes: crawling, kicking, punching, sitting, standing, and walking. The values across the confusion matrix having higher number represent the higher classification accuracy, with activities like crawling, sitting, and standing achieving near-perfect predictions. However, some misclassifications are observed, notably between similar activities such as kicking and punching (78 and 46 cases, respectively) and sitting and standing (56 cases). Walking also shows minor confusion with other activities, likely due to overlapping motion features.

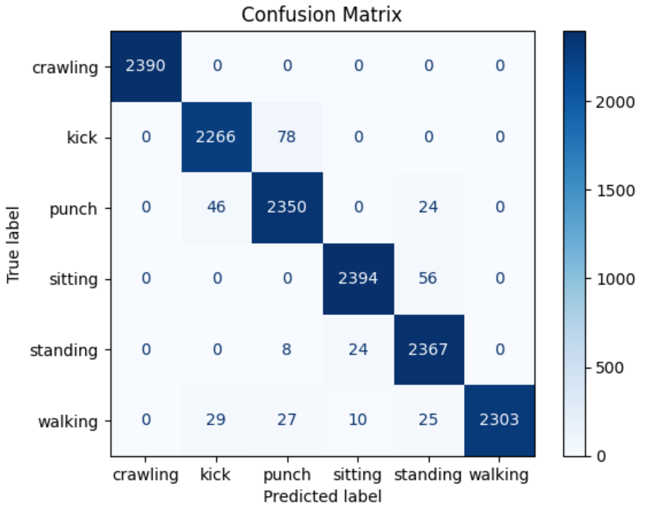


Figure 6. Confusion matrix for each of the activity recognition type.

Overall, the model achieved excellent results, with minimal misclassification errors, highlighting its effectiveness in classifying the activity classes. The model demonstrated an accuracy of 0.9773 on the test set, signifying that about 97.73% of its predictions were accurate. It reflects the model's performance, with precision and recall values close to 1.00 for activities such as crawling and walking. For other activities, the model maintained high scores, achieving 0.97 for kicking, 0.95 for punching, 0.99 for sitting, and 0.96 for standing.

The F1-scores are notably high, ranging between 0.96 and 1.00, indicating a well-balanced performance in terms of precision and recall. Additionally, precision, recall, and F1-score, both at 0.98, underscore the model's reliability and effectiveness in accurately classifying activities across all classes in the test set.

3.3.2. Face Recognition Model Performance

The SVM classifier is used for classifying faces. In this process, faces are first extracted using the MTCNN model. Based on the extracted faces, key points for each individual are identified. These key points are then classified using the SVM classifier to achieve face recognition. Overall accuracy received from this model is 0.99 for classification and with 0.98 for both precision and recall and 0.99 for F1 score, as shown in precision Algorithm (1) below:

Algorithm 1: Integrated Face and Activity Recognition Model.

```
Step1: Load YOLOv8, classifier, MTCNN, FaceNet, and pre-trained models.
Step 2: Define keypoint class for body parts in activity recognition.
Step 3: Extract keypoints for activity recognition from pose estimation results.
Step 4: Generate face embeddings using FaceNet for face recognition.
Step 5: Predict face identity using the classifier model, labeling as "UNKNOWN" if below threshold.
Step 6: Open and process each video frame.
Step 7: Perform activity recognition every frame and predict pose.
Step 8: Detect faces, generate embeddings, and predict identity.
Step 9: Annotate frames with activity and face recognition results.
Step 10: Display the processed frame and exit when 'q' is pressed.
```

4. Results and Analysis

The proposed system is trained, tested and validated on the dataset created using CCTV footage of IR and IP camera, drone and also using benchmark datasets as listed in Table 2. Experimental results are analyzed and compared with existing works to conclude on the novelty of the proposed system and its usage in surveillance environments to detect suspicious activities.

4.1. Integration of Activity and Face Recognition

Integration of these two models is important for the improved surveillance task. This work integrates activity recognition and face recognition models for real-time video processing in surveillance applications. The system uses a YOLOv8-based pose estimation model to identify skeletal key points, which are classified into activities using a custom-trained classifier. Activity recognition is performed on every 10th frame to optimize computational efficiency. For face recognition, the MTCNN model detects faces, and FaceNet generates embedding, which are matched against a pre-trained classifier to identify individuals. Both models are integrated within a shared video processing loop, allowing them to process the same input frames simultaneously. The outputs are visualized on the video, with activity predictions displayed near skeletal keypoints and recognized names shown near face bounding boxes. This unified framework ensures synchronized operation of both tasks, enhancing interpretability. The system demonstrates an efficient multitasking approach, suitable for real-time human behavior monitoring and identity recognition, improving situational awareness in surveillance systems.

Figure 7 illustrates the face recognition output using the drone dataset flying around 5 to 6 feet above the ground. The model was trained to recognize faces present in the database, which includes only two known individuals. All other faces in the dataset are classified as unknown. The model accurately identified the two known individuals, demonstrating its effectiveness. Figures 7 and 8 shows the output of the model, which can detect multiple individuals in a single frame along with their respective activities. Figure 9 presents the other two activities sitting and standing along with the name of the person above the head and activity side of shoulder.

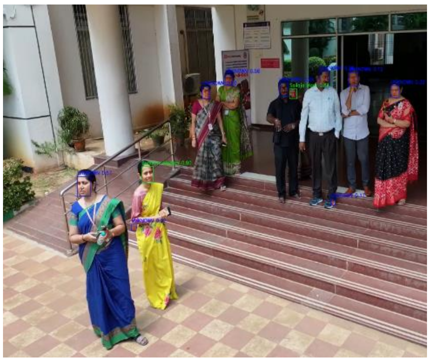


Figure 7. Face recognition model output using drone data.

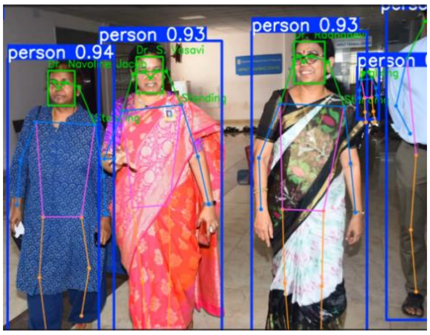


Figure 8. Output of integrated face and activity recognition.

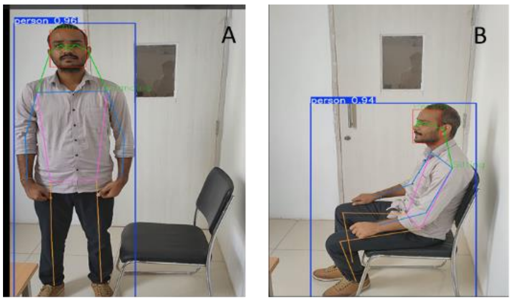


Figure 9. Showing the activity of the two classes such as standing and sitting.

Tables 3 and 4 presents the results of the HAR model w.r.t different activities and also compares with other existing models.

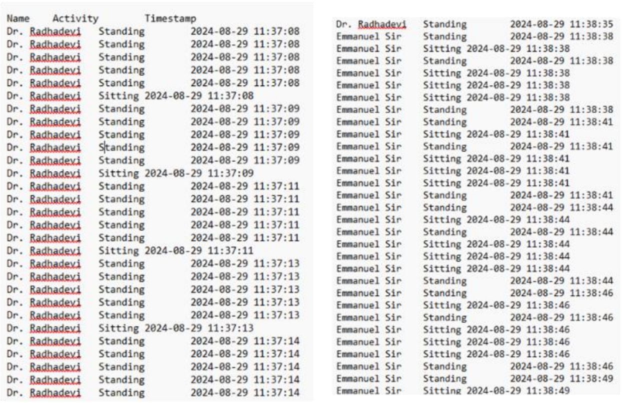
Table 3. Evaluation of HAR model.

Activity	Accuracy
Crawling, kicking and walking	0.97
Punching	0.95
Sitting	0.99
Standing	0.96

Table 4. Comparison of face detection and activity recognition model.

Model	Accuracy
Enhanced MTCNN	0.95
Ensemble model (FaceNet to extract embeddings)	0.82
Ensemble model (Yolo to detect face)	0.95
Ensemble model (SVM for classification)	0.6
[24] for standing activity average of four classifiers [SVM, K-NN, naïve bayes and random forest]	0.857
Ensemble model for Standing activity	0.96
[28] for standing activity average of three classifiers [SVM, k-NN and Random forest]	0.8805
[45] for standing activity using Self-Cascaded Deep Neural Network (SCDNN)	92.27
[45] for s itting and walking using SCDNN	87

Generating a summary for this work is crucial to ensure proper security applications. Using the integrated technique, a summary file of the video is generated. This summary file includes the name of the person performing the activity, along with the corresponding activity. This summary enables effective monitoring of activities and individuals for surveillance purposes. Figure 10 showing the summary report of one video in which multi person are present and performing multi activities at the same time.



```
Name    Activity        Timestamp
Dr. Radhadevi   Standing        2024-08-29 11:37:08
Dr. Radhadevi   Standing        2024-08-29 11:37:08
Dr. Radhadevi   Standing        2024-08-29 11:37:08
Dr. Radhadevi   Standing        2024-08-29 11:37:08
Dr. Radhadevi   Standing        2024-08-29 11:37:08
Dr. Radhadevi   Sitting 2024-08-29 11:37:08
Dr. Radhadevi   Standing        2024-08-29 11:37:09
Dr. Radhadevi   Standing        2024-08-29 11:37:09
Dr. Radhadevi   Standing        2024-08-29 11:37:09
Dr. Radhadevi   Standing        2024-08-29 11:37:09
Dr. Radhadevi   Standing        2024-08-29 11:37:09
Dr. Radhadevi   Sitting 2024-08-29 11:37:09
Dr. Radhadevi   Standing        2024-08-29 11:37:11
Dr. Radhadevi   Standing        2024-08-29 11:37:11
Dr. Radhadevi   Standing        2024-08-29 11:37:11
Dr. Radhadevi   Standing        2024-08-29 11:37:11
Dr. Radhadevi   Standing        2024-08-29 11:37:11
Dr. Radhadevi   Sitting 2024-08-29 11:37:11
Dr. Radhadevi   Standing        2024-08-29 11:37:13
Dr. Radhadevi   Standing        2024-08-29 11:37:13
Dr. Radhadevi   Standing        2024-08-29 11:37:13
Dr. Radhadevi   Standing        2024-08-29 11:37:13
Dr. Radhadevi   Standing        2024-08-29 11:37:13
Dr. Radhadevi   Sitting 2024-08-29 11:37:13
Dr. Radhadevi   Standing        2024-08-29 11:37:14
Dr. Radhadevi   Standing        2024-08-29 11:37:14
Dr. Radhadevi   Standing        2024-08-29 11:37:14
Dr. Radhadevi   Standing        2024-08-29 11:37:14
Dr. Radhadevi   Standing        2024-08-29 11:37:14
```

```
Dr. Radhadevi   Standing        2024-08-29 11:38:35
Emmanuel Sir    Standing        2024-08-29 11:38:38
Emmanuel Sir    Sitting 2024-08-29 11:38:38
Emmanuel Sir    Standing        2024-08-29 11:38:38
Emmanuel Sir    Sitting 2024-08-29 11:38:38
Emmanuel Sir    Sitting 2024-08-29 11:38:38
Emmanuel Sir    Sitting 2024-08-29 11:38:38
Emmanuel Sir    Standing        2024-08-29 11:38:38
Emmanuel Sir    Standing        2024-08-29 11:38:41
Emmanuel Sir    Sitting 2024-08-29 11:38:41
Emmanuel Sir    Standing        2024-08-29 11:38:41
Emmanuel Sir    Sitting 2024-08-29 11:38:41
Emmanuel Sir    Sitting 2024-08-29 11:38:41
Emmanuel Sir    Sitting 2024-08-29 11:38:41
Emmanuel Sir    Standing        2024-08-29 11:38:41
Emmanuel Sir    Standing        2024-08-29 11:38:44
Emmanuel Sir    Sitting 2024-08-29 11:38:44
Emmanuel Sir    Standing        2024-08-29 11:38:44
Emmanuel Sir    Sitting 2024-08-29 11:38:44
Emmanuel Sir    Sitting 2024-08-29 11:38:44
Emmanuel Sir    Sitting 2024-08-29 11:38:44
Emmanuel Sir    Standing        2024-08-29 11:38:44
Emmanuel Sir    Standing        2024-08-29 11:38:46
Emmanuel Sir    Sitting 2024-08-29 11:38:46
Emmanuel Sir    Standing        2024-08-29 11:38:46
Emmanuel Sir    Sitting 2024-08-29 11:38:46
Emmanuel Sir    Sitting 2024-08-29 11:38:46
Emmanuel Sir    Sitting 2024-08-29 11:38:46
Emmanuel Sir    Standing        2024-08-29 11:38:46
Emmanuel Sir    Standing        2024-08-29 11:38:49
Emmanuel Sir    Sitting 2024-08-29 11:38:49
```

Figure 10. Summary file of the processed video.

5. Conclusions and Future Work

In this study, an integrated deep learning and machine learning-based model for face and multi-person human activity recognition is presented, achieving high performance on the input dataset. The proposed model extracts 17 key features using a human skeleton based YOLOv8 model. These key points are subsequently classified into six activity classes. Pose estimation uses ensemble machine learning algorithms to classify the human activities. For face recognition model a pipeline MTCNN is used where FaceNet model extracts the features presented in the faces, SVM classifier for recognition. Such a combined system provided robust framework for face recognition. The integration of a face recognition module with the activity recognition framework enhances the robustness of the system. This combined approach facilitates the development of an alert generation system, which can be effectively utilized for surveillance applications. The integration of these two subsystems provides multiple layers of security by correlating detected activities with recognized individuals, offering proactive surveillance through anomaly detection, personalized alerts, and detailed tracking logs. The system can further be improved for real-time processing under different climatic conditions and can be deployed on edge devices to achieve minimal latency.

Acknowledgment

The authors gratefully acknowledge Advanced Data processing Research Institute (ADRIN), Department of Space, ISRO, for funding this research work Project No: IA-ADRIN-VRSEC-VS-FEB/2024/1. This incredible opportunity made us to contribute integrated face recognition system that addresses key challenges in found in surveillance systems, making significant contributions to the fields of security and human identification.

References

[1] Amraee S., Vafaei A., Jamshidi K., and Adibi P., “Anomaly Detection and Localization in Crowded Scenes Using Connected Component Analysis,” *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 14767-14782, 2018. https://doi.org/10.1007/s11042-017-5061-7

[2] Angelini F., Fu Z., Long Y., Shao L., and Naqvi S., “2D Pose-Based Real-Time Human Action Recognition with Occlusion-Handling,” *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1433-1446, 2020. DOI:10.1109/TMM.2019.2944745

[3] Baek S., Shi Z., Kawade M., and Kim T., “Kinematic-Layout-Aware Random Forests for Depth-Based Action Recognition,” *in Proceedings the of the British Machine Vision*

Conference, London, pp. 1-10, 2017. DOI:10.5244/C.31.13

[4] Bukht T., Rahman H., Shaheen M., Algarni A., Almujally N., and Jalal A., "A Review of Video-Based Human Activity Recognition: Theory, Methods and Applications," *Multimedia Tools and Applications*, vol. 84, pp. 18499-18545, 2024. https://doi.org/10.1007/s11042-024-19711-w

[5] Butt A., Manzoor S., Baig A., Imran A., Ullah I., and Muhammad W., "On-the-Move Heterogeneous Face Recognition in Frequency and Spatial Domain Using Sparse Representation," *PLoS ONE*, vol. 9, no. 10, pp. 1-24, 2024. https://doi.org/10.1371/journal.pone.0308566

[6] Chang Y., Tu Z., Xie W., Luo B., Zhang S., Sui H., and Yuan J., "Video Anomaly Detection with Spatio-Temporal Dissociation," *Pattern Recognition*, vol. 122, no. 4, pp. 108213, 2022. https://doi.org/10.1016/j.patcog.2021.108213

[7] Chong Y. and Tay Y., "Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder," *in Proceedings of the 14th International Symposium*, Sapporo, pp. 189-196, 2017. https://doi.org/10.1007/978-3-319-59081-3_23

[8] Escalera S., Gonzàlez J., Baró X., and Shotton J., "Guest Editors' Introduction to the Special Issue on Multimodal Human Pose Recovery and Behavior Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1489-1491, 2016. DOI:10.1109/TPAMI.2016.2557878

[9] Feichtenhofer C., Fan H., Malik J., and He K., "Slowfast Networks for Video Recognition," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 6202-6211, 2019. DOI:10.1109/ICCV.2019.00630

[10] Feichtenhofer C., Pinz A., and Zisserman A., "Convolutional Two-Stream Network Fusion for Video Action Recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 1933-1941, 2016. DOI:10.1109/CVPR.2016.213

[11] Fu C., Wu X., Hu Y., Huang H., and He R., "DVG-Face: Dual Variational Generation for Heterogeneous Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligenc*e, vol. 44, no. 6, pp. 2938-2952, 2020. DOI:10.1109/TPAMI.2021.3052549

[12] George A. and Marcel S., "Modality Agnostic Heterogeneous Face Recognition with Switch Style Modulators," *in Proceedings of the IEEE International Joint Conference on Biometrics*, Buffalo, pp. 1-10, 2024. DOI:10.1109/IJCB62174.2024.10744437

[13] Hasan M., Choi J., Neumann J., Roy-Chowdhury A., and Davis L., "Learning Temporal Regularity in Video Sequences," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 733-742, 2016. DOI:10.1109/CVPR.2016.86

[14] Kamencay P., Benco M., Mizdos T., and Radil R., "A New Method for Face Recognition Using Convolutional Neural Network," *Advances in Electrical and Electronic Engineering*, vol. 15, no. 4, pp. 663-672, 2017. DOI:10.15598/aeee.v15i4.2389

[15] Kremic E. and Subasi A., "Performance of Random Forest and SVM in Face Recognition," *The International Arab Journal of Information Technology*, vol. 13, no. 2, pp. 287-293, 2016. https://www.ccis2k.org/iajit/PDF/Vol.13,%20No.2/8468.pdf

[16] Li L., Mu X., Li S., and Peng H., "A Review of Face Recognition Technology," *IEEE Access*, vol. 8, pp. 139110-139120, 2020. DOI:10.1109/ACCESS.2020.3011028

[17] Lin J., Gan C., and Han S., "TSM: Temporal Shift Module for Efficient Video Understanding," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 7082-7092, 2019. DOI:10.1109/ICCV.2019.00718

[18] Liu D., Wang N., and Gao X., *Heterogeneous Face Recognition*, Handbook of Face Recognition, Springer, 2023. https://doi.org/10.1007/978-3-031-43567-6_14

[19] Liu H., Ren B., Liu M., and Ding R., "Grouped Temporal Enhancement Module for Human Action Recognition," *in Proceedings of the IEEE International Conference on Image Processing*, Abu Dhabi, pp. 1801-1805, 2020. DOI:10.1109/ICIP40778.2020.9190958

[20] Liu M., Liu H., and Chen C., "Enhanced Skeleton Visualization for View Invariant Human Action Recognition," *Pattern Recognition*, vol. 68, pp. 346-362, 2017. https://doi.org/10.1016/j.patcog.2017.02.030

[21] Liu M., Meng F., Chen C., and Wu S., "Novel Motion Patterns Matter for Practical Skeleton-Based Action Recognition," *in Proceedings of the 37th Conference on Artificial Intelligence*, Washington, pp. 1701-1709, 2023. https://doi.org/10.1609/aaai.v37i2.25258

[22] Liu Q., Zhou Z., Shakya S., Uduthalapally P., Qiao M., and Sung A., "Smartphone Sensor-Based Activity Recognition by Using Machine Learning and Deep Learning Algorithms," *International Journal of Machine Learning and Computing*, vol. 8, no. 2, pp. 121-126, 2018. DOI:10.18178/ijmlc.2018.8.2.674

[23] Naik A. and Naik M., Single Person Violent Activity Dataset, 2022. https://doi.org/10.34740/KAGGLE/DSV/4114209, Last Visited, 2025.

[24] Newaz N. and Hanada E., "A Low-Resolution Infrared Array for Unobtrusive Human Activity

Recognition that Preserves Privacy," *Sensors*, vol. 24, no. 3, pp. 1-16, 2024. https://doi.org/10.3390/s24030926

[25] Rajput S., Bilal M., and Habib A., Human Activity Recognition (HAR-Video Dataset), 2023. https://doi.org/10.34740/KAGGLE/DSV/5722068, Last Visited, 2025.

[26] Ren B., Tang H., Meng F., Runwei D., Torr P., and Sebe N., "Cloth Interactive Transformer for Virtual Try-On," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 4, pp. 1- 20, 2023. https://doi.org/10.1145/3617374

[27] Ren Z., Yuan J., Meng J., and Zhang Z., "Robust Hand Gesture Recognition with Kinect Sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110-1120, 2013. DOI:10.1109/TMM.2013.2246148

[28] Rezaei A., Stevens M., Argha A., Mascheroni A., Puiatti A., and Lovell N., "An Unobtrusive Human Activity Recognition System Using Low Resolution Thermal Sensors, Machine and Deep Learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 115-124, 2023. DOI:10.1109/TBME.2022.3186313

[29] Salehzadeh A., Calitz A., and Greyling J., "Human Activity Recognition Using Deep Electroencephalography Learning," *Biomedical Signal Processing and Control*, vol 62, pp. 102094, 2020. https://doi.org/10.1016/j.bspc.2020.102094

[30] Sultani W., Chen C., and Shah M., "Real-World Anomaly Detection in Surveillance Videos," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 6479-6488, 2018. DOI:10.1109/CVPR.2018.00678

[31] Tarmissi K., Allaaboun H., Abouellil O., Alharbi S., and Soqati S., "Automated Attendance Taking System Using Face Recognition," *in Proceedings of the 21st Learning and Technology Conference*, Jeddah, pp. 19-24, 2024. DOI:10.1109/LT60077.2024.10469452

[32] Thatipelli A., Narayan S., Khan S., Anwer R., Khan F., and Ghanem B., "Spatio-Temporal Relation Modeling for Few-Shot Action Recognition," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, pp. 19958-19967, 2022. DOI:10.1109/CVPR52688.2022.01933

[33] Tolulope O., Funke O., and Adewale O., "Development of an Attendance Management System Using Facial Recognition Technology," *Journal of Engineering Research and Reports*, vol. 26, no. 10, pp. 297-307, 2024. https://doi.org/10.9734/jerr/2024/v26i101307

[34] Tran D., Wang H., Torresani L., Ray J., LeCun Y., and Paluri M., "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, pp. 6450-6459, 2018. DOI:10.1109/CVPR.2018.00675

[35] Tu Z., Liu Y., Zhang Y., Mu Q., and Yuan J., "DTCM: Joint Optimization of Dark Enhancement and Action Recognition in Videos," *IEEE Transactions on Image Processing*, vol. 32, pp. 3507-3520, 2023. DOI:10.1109/TIP.2023.3286254

[36] Verma P., Sah A., and Srivastava R., "Deep Learning-Based Multi-Modal Approach Using RGB and Skeleton Sequences for Human Activity Recognition," *Multimedia Systems*, vol. 26, no. 6, pp. 671-685, 2020. https://doi.org/10.1007/s00530-020-00677-2

[37] Wang L., Xiong Y., Wang Z., Qiao Y., Lin D., Tang X., and Gool L., "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," *in Proceedings of the European Conference on Computer Vision*, Amsterdam, pp. 20-36, 2016. https://doi.org/10.1007/978-3-319-46484-8_2

[38] Wang Y., Kang H., Wu D., Yang W., and Zhang L., "Global and Local Spatio-Temporal Encoder for 3D Human Pose Estimation," *IEEE Transactions on Multimedia*, vol. 26, pp. 4039-4049, 2024. DOI:10.1109/TMM.2023.3321438

[39] Wei J., Jian-qi Z., and Xiang Z., "Face Recognition Method Based on Support Vector Machine and Particle Swarm Optimization," *Expert Systems with Applications: An International Journal*, vol. 38, no. 4, pp. 4390-4393, 2011. https://doi.org/10.1016/j.eswa.2010.09.108

[40] Xu C., Govindarajan L., and Zhang Y., and Cheng L., "Lie-X: Depth Image Based Articulated Object Pose Estimation, Tracking, and Action Recognition on Lie Groups," *International Journal of Computer Vision*, vol. 123, pp. 454-478, 2017. https://doi.org/10.1007/s11263-017-0998-6

[41] Yadav S., Sai S., Gundewar A., Rathore H., Tiwari K., Pandey H., and Mathur M., "CSITime: Privacy-Preserving Human Activity Recognition Using WiFi Channel State Information," *Neural Networks*, vol. 146, pp. 11-21, 2022. https://doi.org/10.1016/j.neunet.2021.11.011

[42] Yadav S., Tiwari K., Pandey H., and Akbar S., "A Review of Multimodal Human Activity Recognition with Special Emphasis on Classification, Applications, Challenges and Future Directions," *Knowledge-Based Systems*, vol. 223, pp. 106970, 2021. https://doi.org/10.1016/j.knosys.2021.106970

[43] Yang F., Wu Y., Sakti S., and Nakamura S., "Make Skeleton-Based Action Recognition Model

Smaller, Faster and Better," *in Proceedings of the 1st ACM International Conference on Multimedia in Asia*, Beijing, pp. 1-6, 2019. https://doi.org/10.1145/3338533.336656

[44] Yang Y., Xian Y., Fu Z., and Naqvi S., "Video Anomaly Detection for Surveillance Based on Effective Frame Area," *in Proceedings of the IEEE 24th International Conference on Information Fusion*, Sun City, pp. 1-5, 2021. DOI:10.23919/FUSION49465.2021.9626932

[45] Yin C., Miao X., Chen J., Jiang H., Chen D., and Tong Y., "Human Activity Recognition with Low-Resolution Infrared Array Sensor Using Semi-Supervised Cross-Domain Neural Networks for Indoor Environment," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11761-11772, 2023. DOI:10.1109/JIOT.2023.3243944

Vasavi Sanikommu working as a Professor CSE and Programme Head Artificial Intelligence and Data science with 27 years of experience. She successfully completed two funded projects and currently seven ongoing funded projects from ISRO, DST and IEEE. She Completed Collaborative research work with Scientists, Industry Experts and Academicians. She visited several countries to present her research articles. She was granted three patents and filed 7 patents.

Sobhana Mummaneni working as an Associate Professor CSE with 18 years of experience. She has currently four ongoing funded projects from ISRO, DST. She is a reviewer for several peer reviewed journals and conferences. She has published 7 patents.

Novaline Jacob working as Deputy Director, ADRIN, Dept of Space, Govt. of India. She has applied various spatial and dynamic modeling techniques, addressing a variety of applications in diverse domains Like Urban, Environment, Disasters, Agriculture, Hydrology, etc. She has established and developed workable methodologies for solving problems in the Field of Land and Water Resources Development, Floods, Landslides, Avalanches, Agriculture Development, Land Degradation and Land Use Change Dynamics studies, etc.

Emmanuel Sanjay Raj K.C is heading Video Analytics Section at ADRIN, Dept. of Space. He received degree in MSc(physics) from Osmania University,1994 and MS (Computer Science), 2002. His research interests include Computer Vision, Intelligent Video Surveillance Systems, Content-based Image retrieval, Digital Watermarking, Visual Cryptography and Steganalysis.

Bhartendra Kumar working as a research associate in the funded project. He completed his M.Tech in Remote Sensing and GIS from the Indian Institute of Remote Sensing (IIRS), Dehradun, in 2024. His research interests include Machine Learning, Computer Vision Engineering and Geoinformatics.

Radha Devi Pullur Variam outstanding Scientist and Director, Advanced Data Processing Research Institute (ADRIN), Department of Space, Hyderabad. She has pioneered in understanding and conceptualizing Satellite Imaging Geometries and Bringing Satellite Photogrammetry Technology as a Prominent Tool for Mapping.