Revolutionary Advances in Computing and Electronics (RACE): An International Journal
Volume 02, Issue 01, pp.15-33, 2026
ISSN (Online): 3107-8540

Received: 24/02/2026, Accepted: 16/04/20026

Anomalous Activity Recognition Using Pose Estimation and Incremental Learning

Nikam Prathmesh Sunil[1], Kummandas Meena[1], Akash Sinha[2]

[1]Center for Artificial Intelligence, Maulana Azad National Institute of Technology, Bhopal, India
[2]Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

25215011106@stu.manit.ac.in, 24215011106@stu.manit.ac.in, akash.sinha@manit.ac.in

ABSTRACT
Automated surveillance is based on the observation of abnormal human behaviors in video streams. In practice, this task remains challenging because abnormal events are inherently rare and difficult to capture in large-scale annotated datasets. Consequently, most supervised approaches depend heavily on labeled instances of anomalies and often exhibit limited generalization in real-world environments, where novel or previously unseen behaviors may arise. To address this weakness, a pose-based anomaly detection system is proposed, which learns patterns of normality human motion. The system does not learn about specific abnormal activities; instead, it learns normal patterns of behaviour and identifies significant deviations as anomalies. Video frames are processed to extract human skeletal key points using the YOLO-Pose model, which predicts 17 body joints per person. Keypoints are normalized with respect to the bounding box coordinates to achieve scale and position invariance. Derived features of spatial and temporal motion the skeleton representation record the movement dynamics and the body posture. The distribution of normal poses is reflected in an Incremental One-Class Support Vector Machine (SGDOneClassSVM), which is trained only on normal samples. The experiments on the ShanghaiTech Campus dataset show that the proposed method achieves a detection accuracy of 91.0% and operates at about 21 frames per second.

Keywords: *Human Action Recognition, Anomaly Detection, YOLO Pose, One-Class SVM, Incremental Learning, Real-time Surveillance, Skeleton-based Analysis*

1 Introduction

Understanding human activities is a fundamental component of intelligent surveillance systems. Surveillance systems are widely deployed in public and private environments, including schools, offices, airports, and retail spaces. Therefore, automated systems are required to analyze video data and detect abnormal activities. Traditional surveillance relies on manual monitoring at the same time. This approach is inefficient because people can get tired, miss important details, or react too slowly.

Significant progress has been made with intelligence and computer vision. Now proposed a system that help, or even replace, people watching cameras. These systems try to find activities by looking at patterns in video. For example, some people have written about how to find activities in video [1]. Making systems that can do this job well is still very hard. Things

*Corresponding author: Nikam Prathmesh Sunil, Center for Artificial Intelligence, Maulana Azad National Institute of Technology Bhopal, India (25215011106@stu.manit.ac.in)

that are not normal do not happen often and we cannot predict when they will happen. So, it is hard to get a lot of examples of behavior. That is why many people try to learn what normal behaviour looks like and then identify anything that differs from that as a potential problem [2]. Surveillance systems are very important. They need to operate effectively. Surveillance systems need to be smart and find unusual activities.

Analyzing how people move is an effective way to understand human activities in video. Instead of analyzing the entire image, skeletal keypoints can be used to represent the positions of body joints. Modern pose estimation methods, such as OpenPose, can accurately detect body joints in real time [3]. This representation focuses on body movement and posture while being less sensitive to background variations and lighting conditions. Because of this, pose-based representations are useful for surveillance applications [4]. Another challenge in surveillance systems is that normal behavior can change over time. Activities that are considered normal in one environment or time period may differ in another. Therefore, anomaly detection systems must be able to adapt without requiring complete retraining. Incremental learning techniques address this issue by allowing models to update continuously as new data becomes available [5].

In this paper, we are offering a system that integrates human poses with incremental learning. The system then removes skeletal keypoints from video frames using YOLO-Pose [6]. These keypoints are transformed into feature vectors representing posture and motion patterns. A one-class Incremental Support Vector Machine is then used to learn the distribution of normal activities. The system does not learn predefined abnormal behaviour; it only detects anomalies by detecting deviations from previously learned normal patterns. The proposed system is expected to be efficient and applicable to the real-time applications. Although the system uses skeletal representations in place of raw image data, this helps conserve privacy and makes it resistant to environmental changes. Moreover, incremental learning mechanism enables the model to adhere to alterations in the human behavior without necessarily undergoing complete retraining.

The anomaly-detection system for human activity can be highly useful in surveillance operations such as the one proposed here. Speedy identification of the abnormal human behavior is significant to most intelligent monitoring systems. The suggested framework addresses this issue by integrating elements of pose estimation and incremental learning to develop an adaptive, efficient anomaly detection system.

2 Related Work

This section reviews previous studies on pose estimation, anomaly detection, and incremental learning methods applied in surveillance systems. These studies demonstrate achievements in these spheres while revealing research gaps that motivate the proposed work.

The process of human pose estimation has evolved from primitive manual methods to current deep learning algorithms. Initial attempts, such as pictorial structures and deformable part models, were adopted for modelling the human body form [7]. These techniques were computationally effective but showed poor performance on occlusions as well as poses that tend to occur in the real world. The introduction of deep convolutional neural networks witnessed a drastic improvement in the performance of pose estimation problems. Toshev and Szegedy proposed DeepPose and pose estimation was formulated as a regression problem

which can be solved using deep neural networks [8]. Their study was able to prove that deep learning techniques were capable of projecting correct keypoint positioning. Newell et al. later proposed stacked hourglass networks, which learn the occurrence of associations between body parts by repetitively processing features in bottom-up and top-down information streams
[9].

Multi-person pose estimation systems took the field further. OpenPose identifies poses of multiple people in real-time using Part Affinity Fields to identify recognisable body parts to individuals in the environment [3]. Although such a technique is able to do well in crowd scenes, it involves several processing stages and is computationally very intensive. There are other approaches such as AlphaPose offering better pose estimation with better network designs and significant suppression policies [10]. HRNet keeps high-resolution representations throughout the network, which helps to localise keypoints better [11]. While such techniques have high accuracy, they generally require substantial computational resources.

Recent development has focussed on pose estimation for real-time applications, making the process more efficient. YOLO-Pose is a single-stage architecture that integrates object detection and keypoint estimation to achieve faster and efficient inference while achieving competitive results [6]. Due to the light build, it can be used for edge devices and real-time surveillance systems.

Anomaly detection in the surveillance domain has been well studied. Early approaches were to use handcrafted features, in combination with statistical models or one-class classifiers. For example Mahadevan et al. simulated normal crowd behaviour using mixtures of dynamic textures and separated abnormalities as deviations from learned normal behaviour [1]. While working well in certain situations, such methods tend not to generalise to different situations.

More recent studies have used deep learning for anomaly detection. Hasan et al. introduced a spatio-temporal autoencoder network that learns to reconstruct normal video sequences, and suboptimal reconstruction errors reveal abnormal occurrences [12]. Variational autoencoders are an extension of this work that add probabilistic latent representations that give estimates for the uncertainty in the reconstruction in addition to the reconstruction score. Generative adversarial networks (GANs) have also been used in anomaly detection. Ravanbakhsh et al. trained generative adversarial network based models on normal video data and used the output of discriminator as the anomaly score [13]. However, there is instability in training GAN-based approaches and also the need to optimise the hyperparameters carefully.

Skeleton-based anomaly detection provides privacy-preserving benefits in that it analyses body movement and not visual appearance. Morais et al. showed that it is possible to analyse the information carried by surveillance videos for human activity based on skeletal motion patterns [4]. Similarly, Luo et al. chose to use graph convolutional networks by representing the human skeleton as a graph with body joints as the nodes and the relationships between body joints as their edges to capture spatial-temporal dependencies [14]. These studies show that pose-based representations can be used to represent human behaviour in a privacy-preserving way.

One-class classification is another popular method of anomaly detection. These methods model the normal distribution of data, and the abnormalities are identified as exceptions.

One-Class Support Vector Machine is a popular algorithm to learn the boundary of normal data in feature space [2]. To improve anomaly detection of high-dimensional spaces, Erfani et al. combined deep belief networks and One-Class SVM to better the results when studying high-dimensional feature spaces [15].

A fundamental problem with surveillance systems is the dynamic nature of the real world. Lighting conditions, frequency of people and behavioural patterns of humans can vary over time. Incremental learning provides a way out of this problem by allowing models to update themselves continuously as new information becomes available. Ross et al. demonstrated incremental visual tracking by updating appearance models using low-dimensional subspace representations to allow models to adapt to new observations [16].

Recently, anomaly detection approaches have been investigated that explore deep incremental learning. Tavakoli et al. used incremental learning based on denoising autoencoders and showed a decrease in false positives due to the continuous adaptation of the model to previously unobserved variations in normal behaviour [**?**]. Pang et al. used deep incremental approaches with the use of memory replay to handle catastrophic forgetting, which improved the balance between new information and previous knowledge [18]. Hybrid approaches use deep feature extraction and incremental classifiers. Liu et al. used pre-trained convolutional neural networks for feature extraction and incremental SVM classifiers to realise real-time processing with adaptability to changing environments [**?**].

Despite these developments, little research has been conducted on incremental learning in the specific context of pose-based anomaly detection. The current state-of-the-art techniques either process video frames directly or need to retrain in batches when new data is available. The coupling of efficient pose estimation and progressive one-class classification presents an interesting research direction that has not been explored much, especially for privacy-preserving surveillance systems.

Motivated by these observations, this work proposes a framework that combines efficient pose

estimation using YOLO-Pose with incremental One-Class SVM classification. The proposed approach extracts skeletal representations from video frames and models normal human behavior using an adaptive anomaly detection framework. By integrating real-time pose estimation with incremental learning, the system provides an efficient, privacy-preserving, and adaptive solution for anomaly detection in surveillance environments.

3 Research Methodology

The proposed human action anomaly detection system is presented in the following section of this paper, along with a reference to its methodology. The architecture is pipeline based, which uses the combination of computer vision and incremental machine learning. The design is made in a manner that is reproducible and systematically experimental, and evaluates the proposed approach rigorously as shown in Figure 1.

3.1 Research Design

One of its methods is a data-based approach where skeletal key points are obtained from video frames to represent human motion. The goal is to represent the normal human behavior

and determine abnormal behavior as deviation. Common controlled conditions, such as classrooms and corridors, are a test of the system's operation under realistic conditions.

3.2 Dataset Preparation

All experiments are done over ShanghaiTech Campus dataset that is composed of a collection of video sequences shot at thirteen different scenes under different conditions of illumination and camera positions. Because the method of classification is one-class, the model is trained with normal human behavior samples only. The estimation of poses is done with YOLO-Pose that identifies 17 body parts of the person.
The keypoints are encoded as the spatial coordinates of a keypoint and this creates a 34-dimensional pose feature vector on normalizing.

The dataset is separated into training and experimental datasets. The training program includes normal human activities such as walking, sitting, standing, writing, bending, and basic gestures. Originally, training is done on about 2,500 samples, as shown in Table 1.

Both normal and anomalous samples are present in the test set, which will comprise 500 normal and 400 anomalous samples, for a total of 900. Running, falling, excessive bending and abnormal postures are all anomalous activities. In the evaluation, about 15000 video frames are handled.

Table 1: Dataset Composition of ShanghaiTech Subset

Parameter	Value
Initial Training Samples	2,500 – 3,000
Incremental Training Samples	1,500 – 2,000
Total Training Samples	4,000 – 5,000
Normal Test Samples	500
Anomalous Test Samples	400
Total Test Samples	900
Total Frames Processed	~15,000

3.3 Feature Engineering

Pose-based analysis involves representations that remain invariant to the location of the subject in the image or the scale of the individual relative to the camera. To achieve this, raw keypoint coordinates are normalized using the bounding box surrounding each detected person.

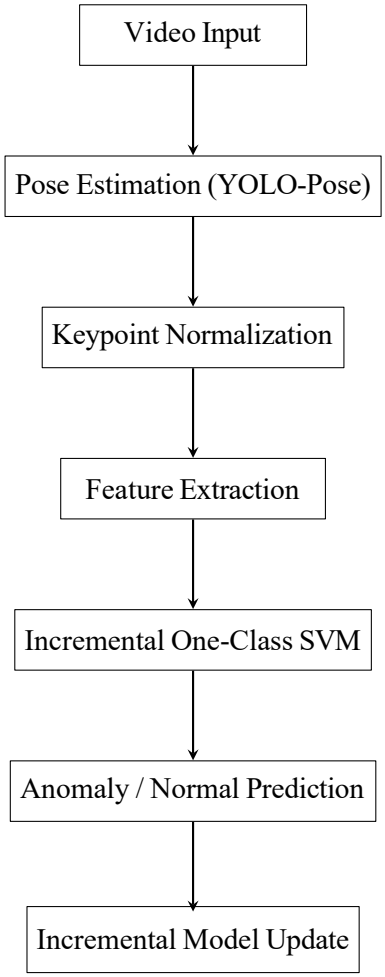

Figure 1: Pipeline of the Proposed Human Action Anomaly Detection System

Let $p_i = (x_i, y_i)$ denote a keypoint coordinate and $(x_{min}, y_{min}, x_{max}, y_{max})$ denote the bounding box coordinates. The normalized coordinates are computed as:

$$\hat{x}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$$\hat{y}_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \quad (2)$$

Such normalization maps all keypoints into the unit square $[0, 1] \times [0, 1]$, removing dependence on absolute position and scale.

For static pose analysis, the normalized keypoints are aggregated into a 34-dimensional feature vector:

$$F_{static} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \ldots, \hat{x}_{17}, \hat{y}_{17}] \in \mathrm{R}^{34} \quad (3)$$

To capture motion dynamics, temporal features are extracted from consecutive frames. The velocity of each keypoint is calculated as:

$$v \quad \begin{matrix} \mathrm{S} \\ \mathrm{i} \end{matrix} \quad \begin{matrix} \text{milarly,} \\ \text{acceleratio} \end{matrix} \quad \begin{matrix} \mathrm{n} \\ \text{is} \end{matrix}$$

computed as:

$$a_{j,t} = \frac{p_{j,t} - p_{j,t-1}}{\Delta t} \tag{4}$$

$$_{j,t} = \frac{v_{j,t} - v_{j,t-1}}{\Delta t} \tag{5}$$

The final temporal representation is obtained by concatenating pose, velocity, and acceleration features:

$$Ftemporal = [pt, vt, at] \in R^{102} \quad (6)$$

Additional geometric parameters such as joint angles and normalized limb lengths are also calculated to capture spatial relationships between body joints. These features help the model better differentiate between various human behaviors and remain invariant to camera angle and scale.

3.4 Experimental Configuration

The system is evaluated under a controlled experimental setup designed to simulate real-world surveil- lance conditions.

The model is initially trained using 2,500–3,000 normal samples. During deployment, newly observed normal samples are stored in a buffer. Once the buffer reaches a predefined size of 100 samples, the model is updated incrementally.

A confidence threshold of 0.65 is used to ensure that only reliable normal samples are included. Frame sampling is performed at a rate of one frame every three frames to reduce computational overhead.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN). \quad (7)$$

Precision indicates the rate of occurrence of detected anomaly alerts that are correct:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

Recall is the percentage of actual anomalies identified by the system:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

F1-score is a balanced score of precision and recall that is defined as a harmonic mean of precision and recall:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (10)$$

Other than classification accuracy, detection latency was also used to assess temporal performance. Detection latency is defined as a time difference between the event of occurrence of an anomalous event and the time when the system alerts. It is computed as:

$$\text{Detection Latency} = t_{\text{alert}} - t_{\text{occurrence}}. \quad (11)$$

Real time response was also tested in terms of processing speed that was measured as the number of frames per second (FPS). This measure of the system is the number of video frames that are handled by the system in one second and is described as:

$$\text{FPS} = \text{Frames Processed} / \text{Processing Time (s)} \quad (12)$$

Time-series cross-validation using five consecutive splits was done to recreate realistic conditions of operations. In this method, previous video sequences were employed in the training and subsequent sequences were employed during testing. The strategy also maintains the time sequence of the data and does not allow the model to access future information during training, therefore, avoiding look-ahead bias.

3.5 Incremental Learning

Incremental learning enables continuous model adaptation without complete retraining.

The model is initially trained on a dataset D_0 consisting of normal samples. During real-time operation, new samples classified as normal with high confidence are stored in a buffer B.

Once the buffer reaches a predefined size $U = 100$, the model is updated using a partial fitting process. After updating, the buffer is cleared for subsequent data collection.

This approach allows the system to progressively adapt to evolving patterns of normal behavior while maintaining computational efficiency.

3.6 Theoretical Foundations

This section states the major concepts of the proposed human action anomaly detection system. The methodology is a combination of computer vision, geometric modeling of human motion and statistical learning so as to model human movement and detect abnormal patterns of activities.

3.6.1 Pose Representation

Human pose estimation is the task of finding the anatomical keypoints of the human body in a given image or video frame. The current pose estimation models can detect and regress keypoints to objects using a single forward pass. This work uses the YOLO-Pose architecture to extract skeletal keypoints in each video frame (Maiji 2022).

The process of pose estimation may be developed as a transformation between an input image and a set of individuals detected and their corresponding keypoints:

$$\Psi(I) = \{B_i, K_i\}_{i=1}^{N} \tag{13}$$

In which, I is the input image, B_i is the bounding box of the ith identified person, K_i is the skeletal keypoints of the identified person and N is the number of persons identified in the frame.

Each keypoint is described by the spatial coordinates and confidence value:

$$P_{ij} = (x_{ij}, y_{ij}, c_{ij}) \tag{14}$$

and the values of x_{ij} and y_{ij} represent the position of the jth keypoint of the ith individual, and c_j is the confidence of the keypoint detection.

3.6.2 Feature Representation

Raw keypoint coordinates are transformed to normal values with respect to the bounding box of each person detected to be camera position and scale-invariant. Assuming a keypoint of the form (x_{ij}, y_{ij}) and the coordinates of a bounding box of the form $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, the normalized coordinates can be calculated as:

$$\hat{x}_{ij} = (x_{ij} - x_{\min})/(x_{\max} - x_{\min}) \tag{15}$$

$$\hat{y}_{ij} = (y_{ij} - y_{\min})/(y_{\max} - y_{\min}). \tag{16}$$

This normalization transforms all the keypoints into the unit square where they fall in $[0, 1] \times [0, 1]$, without relying on absolute position and scale, yet geometric connections between joints are still obtained.

The keypoints are normalized and concatenated into a static pose feature vector:

$$f_{\text{static}} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \ldots, \hat{x}_{17}, \hat{y}_{17}] \in \mathrm{R}^{34}. \tag{17}$$

This is a space-time representation of the skeletal arrangement of a human body at a particular moment.

This section discusses the temporal motion representation.

One of the significant aspects in interpreting human activities based on video streams is temporal motion representation.

Human pose information in a single frame is usually analyzed and it is believed that the information is adequate in determining human activities. Nonetheless, motion patterns evolve as human activities in many cases. So there is the need to include the time information through monitoring the variation of body keypoints among successive frames.

The motion dynamics of each body keypoint are calculated by calculating the difference between two consecutive frames. The velocity is defined as

$$v_j^t = \frac{p_j^t - p_j^{t-1}}{\Delta t} \tag{18}$$

in which p_j^t is the position of the jth keypoint at time t, and Δt is the time per frame between pairs of successive frames. Besides velocity, acceleration is also calculated to explain the change in

motion over time.
Acceleration is a rate of change of velocity between frames, and is given by:

$$a_j^t = \frac{v_j^t - v_j^{t-1}}{\Delta t} \tag{19}$$

The spatial pose data, velocity and acceleration are then fused together into one feature representation so as to describe the overall movement properties at a particular time step:

$$f_t = \{p^t, v^t, a^t\} \tag{20}$$

This representation combines both the spatial organization of the human body and the dynamic movement properties; this means that the model is better suited to capture movement patterns and differentiate among various human activities.

3.6.3 One-Class Classification

When working on it, we consider the identification of unusual activities as a one-class classification problem. In essence, we would like the model to be trained on what normal human beings behave like and anything that appears to be very unlike this normal stuff can be labeled as unnatural.

To do this we have employed an algorithm known as One-Class Support Vector Machine (OC-SVM) that was invented by Scholkopf and his colleagues [2]. The effect of this is to learn some kind of a boundary around all the normal data points in our feature space. Suppose we have a set of normal pose vectors denoted by the set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. The mathematics behind it is:

$$\min_{\mathbf{w},\boldsymbol{\xi},\rho} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho \tag{21}$$

subject to the constraint that:

$$\mathbf{w} \cdot \Phi(\mathbf{x}_i) \geq \rho - \xi_i \quad \xi_i \geq 0, \quad \forall i \tag{22}$$

The meaning of these equations is that we are putting our data into a new space by using a kernel function Φ, with slack variables ξ_i which allow for not hitting exactly the correct boundary, and ν which acts as a tradeoff between how many errors we are willing to tolerate in our training and how complex the model we are building would be.

To test whether a new sample is normal or not, we use this decision function:

$$f(\mathbf{x}) = -\operatorname{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x}) - \rho) \tag{23}$$

When we get a negative value of $f(\mathbf{x})$ then we can say that the sample is unusual or anomalous.

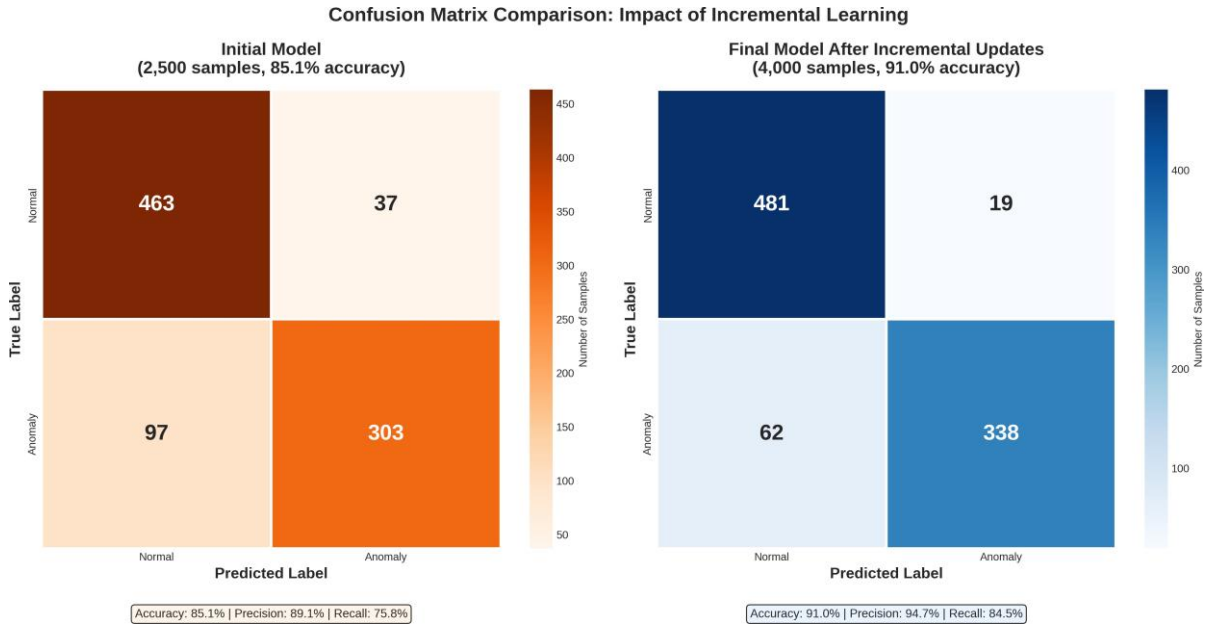

Figure 2: Comparison of confusion matrices before and after incremental learning

3.6.4 Geometric Feature Representation

In addition to pose and motion features, geometric features derived from the skeletal structure are used to improve the discrimination between different human activities. Joint angles provide a viewpoint-invariant description of body configuration.

The angle formed by three joints i, j, and k can be computed as

$$\theta_{ijk} = \arccos \frac{(\mathbf{p}_i - \mathbf{p}_j) \cdot (\mathbf{p}_k - \mathbf{p}_j)}{\|\mathbf{p}_i - \mathbf{p}_j\| \|\mathbf{p}_k - \mathbf{p}_j\|} \tag{24}$$

Furthermore, limb lengths can be normalized by the bounding box height in order to obtain scale- invariant measurements:

$$l_{ij} = \frac{\|\mathbf{p}_i - \mathbf{p}_j\|}{\text{height}_{\text{bbox}}} \tag{25}$$

These geometric descriptors complement the pose and motion features by capturing structural relationships among body joints, thereby improving the robustness of the activity representation.

4 Results and Discussion

This section presents the experimental results obtained from the proposed human action anomaly detection system and analyses the effectiveness of combining pose estimation with incremental One-Class SVM for detecting abnormal activities.

4.1 Confusion Matrix Analysis

Figure 2 compares the confusion matrices for the initial and final models after incremental learning updates. The initial model correctly classified 463 normal samples and detected 303 anomalous samples, with 37 false positives and 97 false negatives. After incremental updates, the model achieved 481 correct normal classifications and 338 correct anomaly detections, while reducing false positives to 19 and false negatives to 62. Based on these values, the main performance metrics were calculated.

As described below, performance improvements are achieved through incremental learning. Table 2 indicates the reduction in false positives as incremental learning updates were applied. As more normal behavior samples were observed, the decision boundary became closer to the ideal, reducing the number of false alarms generated by the model.

Table 2: Progressive False Positive Reduction with Incremental Learning

Update Stage	Cumulative Samples	False Positives (per 200 frames)	Reduction (%)
Initial Model	2,500	18	—
After First Update	3,000	12	33.3
After Second Update	3,500	8	55.6
After Third Update	4,000	5	72.2

Performance improvements are observed with the introduction of incremental learning. Table 2 shows a progressive reduction in false positives as additional normal samples are incorporated into the model.

As the model is exposed to more normal behavior patterns, the decision boundary becomes more refined, resulting in fewer false alarms. This indicates that incremental learning improves the model's ability to distinguish between normal and anomalous activities.

The system operates at approximately 21 frames per second, enabling real-time performance. The average detection latency is 159 ms, with a range of 142–178 ms across

activity types.

Detection performance varies across anomaly types. Running activities achieve approximately 78% detection accuracy due to distinct motion patterns. Falling actions achieve around 77.5% accuracy, particularly when sudden posture changes occur. However, more subtle activities, such as excessive bending or leaning, achieve lower accuracy (72–74%) because they are similar to normal movements.

To understand what causes our system to fail in certain situations, we reviewed instances where it was faulty. Table 3 highlights the primary causes of detection error occurrence.

Table 3: Failure Analysis of Detection Errors

Reason for failure	Percentage
Similar to typical movements	34%
Occlusion leading to poor pose estimation	28%
Changed lighting that influences keypoint detection	18%
Multiple individuals intersecting in frame	12%
Sudden camera movement	8%

Table 4: Failure Mode Analysis

Failure Reason/Cause of Failure	Percentage of Failures
Occlusion (Person Overlap)	35%
Low Lighting Conditions	28%
Fast Transitions / Motion Blur	22%
Action Ambiguity	15%

Most of the failures in detection were as a result of occlusion whereby overlapped people led to incomplete or absent keypoint detection. The confidence of pose estimation was also harsh due to low lighting conditions and rapid movements sometimes created motion blur which lowered the accuracy of keypoint localization.

4.2 Comparison with Current Approaches

Recent technology has been introduced in print media and journalism. We compared our performance in terms of our methodology to a slight portion of other currently existing approaches to detect anomalies. Table 5 compares them all against one another.

Table 5: Comparison to Existing Methods

Method	Accuracy (%)	FPS	Level of Privacy
OpenPose + LSTM	88.5	8	Medium
3D CNN	91.2	5	Low
Autoencoder Reconstruction	87.3	12	Medium

Proposed Method (Initial)	85.1	21	High
Proposed Method (Incremental)	91.0	21	High

By examining the outcome, it is possible to infer that our approach performs quite decently in comparison with other deep learning methods. We are almost equal to them and even better when we added incremental learning component. And the best part of it is that we are much faster; we can run about 21 frames per second, which is very fast compared to most other methods. The same is another advantage of our system that it secures the privacy of people better. The fact that we make use of skeleton poses rather than real video frames means that we are not storing or processing actual images of people.

4.3 Discussion of Key Findings

The results indicate that the pose-based anomaly detection system is effective in identifying abnormal human behavior while preserving privacy. Incremental learning significantly improves long-term performance. As the model is exposed to more normal behavior patterns, it continuously refines its definition of normality, thereby reducing false positives.

The use of skeletal representations reduces computational complexity compared to raw image-based methods. The combination of pose, velocity, and acceleration features provides sufficient information to model human motion dynamics. Additionally, processing at a reduced frame rate (one frame every three frames) enables improved computational efficiency, highlighting a trade-off between speed and accuracy.

4.4 Limitations

Despite its effectiveness, the proposed system has several limitations.

First, the one-class classification approach relies solely on normal data. As a result, the model does not explicitly learn representations of anomalous activities, which may lead to missed detections or false positives in certain cases.

Second, skeletal representations limit contextual understanding. Interactions with objects and environmental changes are not captured, which may restrict the detection of certain types of anomalies.

Finally, system performance depends on the accuracy of pose estimation. Adverse conditions such as poor lighting and high crowd density can lead to errors in keypoint detection, which in turn affect anomaly detection performance.

4.5 Environment-Specific Performance

Our system was also tested in various kinds of places to determine its ability to cope with various patterns of activities. What we found is illustrated in Table 6.

Table 6: Performance Metrics by Deployment Environment

Environment	Accuracy (%)	False Positives per Hour
Classroom	93.2	8
Corridor	89.5	15
Laboratory	91.8	10

Library	94.1	5

Based on these findings, it is possible to notice that the system performs more effectively in those areas where movements of people are more predictable. In libraries and classrooms we had greater accuracy since depending on the type of position of individuals, many sit, stand or walk in fairly predictable modes. But the corridors were more complicated - there were more people around, constantly passing by, and it was more difficult to see people properly with the pose detector. All that resulted in additional errors and false alarms in corridors.

Thus on the whole, it appears that this system does indeed work reasonably well in real time, when it comes to identifying some unusual activities with little more than skeleton movements. Through pose data and gradually updating One-Class SVM learning, we could identify anomalies well without millions of computers processing the data or invasion of privacy of people. And the best thing is that the system is constantly being improved as time goes by - because it gets exposed to more normal activities, it becomes progressively more accurate in its vision of what it deems normal and it only reduces commonality of false alarms and makes it more accurate after a long time.

5 Conclusion

In this paper, an anomaly detection system based on pose estimation and incremental one-class classification has been introduced as a real-time system. Skeletal keypoints detected by YOLO-Pose depict human actions and can be used to preserve individuals' privacy while simultaneously capturing significant motion patterns to detect abnormalities in surveillance settings. The computational efficiency of the system, which is viewpoint independent and not sensitive to changes in the environment such as lighting, and background distortion, is because it uses skeletal representations but not raw pixel data. The evaluation on the ShanghaiTech Campus dataset shows that the proposed framework can provide trustworthy detection with 91.0% accuracy and real-time processing at 21 frames per second. The incremental approach of learning using SGDOneClassSVM is effective in reducing false alarms because the decision boundary is repeatedly narrowed following a successive normal pattern of behaviour. The statistical analysis shows that the difference between accuracy (85.1% and 91.0%) with the help of incremental updates is statistically significant, proving the success of a mechanism of adaptive learning

Pose-based representations offer a privacy-preserving alternative to traditional video analysis and capture motion dynamics. The system allows complete representation of a person's activities through the combination of spatial (normalised keypoint positions) and temporal (velocity and acceleration) features. In addition, incremental learning is critical for real-world deployment due to its ability to add new adaptable behaviour patterns to the system and to minimise false positives when a model is updated, from 37 to 19. Despite these benefits, there are several constraints. Pose representations do not convey the context of the object interactions or variations at the scene level which can limit the capabilities of the system to identify some kinds of anomalies. Moreover, problematic situations such as bad lighting, occlusion and motion blur may also impair performance, by adding detection errors in keypoint detections, which account for a fraction of the observed failures.

The proposed system finds practical uses in healthcare surveillance, surveillance in the field of public safety and industrial surveillance systems wherein real-time detection of

anomalies is necessary. Future research will examine how to integrate temporal deep learning networks, such as LSTM or GRU networks, and begin exploring graph convolutional networks (GCNs) to model the skeleton, integrating multimodal data such as optical flow and object interactions. It might be further extended through deployment to edge computing platforms and the exploration of self-supervised learning techniques to achieve greater scalability and greater reliance on labelled training data.

Funding Source

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, 2010, pp. 1975–1981, doi: 10.1109/CVPR.2010.5539872.

[2] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001, doi: 10.1162/089976601750264965.

[3] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1,
pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[4] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 11996–12004, doi: 10.1109/CVPR.2019.01227.

[5] V. Losing, B. Hammer, and H. Wersing, "Incremental on-line learning: A review and comparison of state of the art algorithms," *Neurocomputing*, vol. 275, pp. 1261–1274, 2018, doi: 10.1016/j.neucom.2017.06.070.

[6] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for multi-person pose estimation using object keypoint similarity," in *Proceedings of the IEEE Conference on Computer Vi- sion and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 2022, pp. 2571–2580, doi: 10.1109/CVPRW56347.2022.00288.

[7] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005, doi: 10.1023/B:VISI.0000042934.15159.49.

[8] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), Columbus, OH, USA, 2014, pp. 1653–1660, doi: 10.1109/CVPR.2014.214.

[9] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 483–499.

[10] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2334–2343, doi: 10.1109/ICCV.2017.256.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 5693–5703, doi: 10.1109/CVPR.2019.00584.

[12] M. Hasan, J. Choi, J. Neumann, A. Roy-Chowdhury, and L. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 733–742, doi: 10.1109/CVPR.2016.86.

[13] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 2017, pp. 189–197, doi: 10.1109/WACV.2017.28.

[14] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, Jul. 2021, doi: 10.1016/j.neucom.2020.07.131.

[15] S. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, Oct. 2016, doi: 10.1016/j.patcog.2016.03.028.

[16] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125–141, May 2008, doi: 10.1007/s11263-007-0075-7.

[17] G. Zhou, K. Sohn, and H. Lee, "Online incremental feature learning with denoising autoencoders," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, La Palma, Canary Islands, 2012, pp. 1453–1461, url: https://proceedings.mlr.press/v22/zhou12b.html.

[18] G. Pang, C. Shen, and A. van den Hengel, "Deep learning for anomaly detection: a review," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 7890–7899, doi: 10.1109/CVPR46437.2021.00780.

[19] A. Hussain, W. Ullah, N. Khan, Z. A. Khan, H. Yar, and S. W. Baik, "Class-incremental learning network for real-time anomaly recognition in surveillance environments,"

Pattern Recognition, vol. 170, p. 112064, Feb. 2026, doi: 10.1016/j.patcog.2025.112064.

[20] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection – A new baseline,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6536–6545, doi: 10.1109/CVPR.2018.00685.