scientific reports

OPEN

Fine-grained temporal–spatial cues for theft recognition in surveillance videos

Mohd. Aquib Ansari[1], Arvind Mewada[2], Ambrish Kumar[2], Ruchi Jayaswal[3], Amrendra Singh Yadav[4], Lalit Kumar[5] & Deepika Bansal[6]✉

Surveillance systems play a crucial role in detecting suspicious human activities, including attacks, violence, and abductions, in public spaces. This study presents a human intervention-free, hybrid framework that utilizes deep neural networks for real-time theft activity recognition. The proposed methodology employs a dual stream fusion network, combining appearance and motion features, to accurately identify theft actions. Specifically, a modified InceptionV3 model extracts relevant body pose features through keypoint transfer, feeding two separate deep neural network pipelines for appearance and motion analysis. Long-Short-Term Memory network then models temporal relationships between the extracted features across consecutive frames. The novelty of this research lies in the proposed dual-stream fusion architecture, which aims to capture fine-grained temporal and spatial cues for theft detection. A new lab-lifting dataset has also been developed to reflect subtle theft behaviors in academic settings. The framework's performance is evaluated on a dataset comprising normal and theft activities. The results demonstrate a recognition accuracy of 91.86% , surpassing that of other methods.

Keywords Surveillance Systems, Theft Detection, Deep Neural Networks (DNN), Dual-Stream Fusion, Deep Inceptionv3 Model, Long Short-Term Memory (LSTM), Human Activity Recognition

Surveillance systems play a crucial role in maintaining public safety by monitoring and detecting suspicious human activities, including violence, theft, and other criminal behaviors[1]. Traditional closed-circuit television (CCTV) surveillance systems are widely deployed in public places such as schools, hospitals, and commercial spaces[2]. Common datasets used for abnormality detection tasks include KTH, Weizmann, UCSD, and Boss, encompassing actions such as walking, running, and fighting. Some studies have employed specifically curated datasets for violence detection[3]. A significant challenge in human activity recognition is the lack of comprehensive datasets for real-world behavior recognition. However, these systems depend highly on human operators for observation and detection, which introduces significant challenges, including fatigue, oversight, and delayed response[4,5]. Moreover, surveillance's increasing complexity and scale demand automated solutions to improve accuracy, reliability, and security.

Related work

Integrating Internet of Things (IoT) devices with artificial intelligence (AI) can present a promising approach to tackle this limitation. In real time, these AI-enabled surveillance systems can autonomously identify abnormal behaviors, such as fights, shoplifting, and violent acts, facilitating proactive intervention[6]. Various methodologies for action recognition have been explored, ranging from statistical techniques to probabilistic and machine learning (ML)-based approaches[6–9]. These techniques generally consist of four key components: data collection, feature extraction, modeling, and activity classification. Conventional feature extraction methods include histograms of oriented gradients (HOG)[10], motion boundary histograms (MBH)[11], space-time interest points[12], and optical flow histograms[13]. These features are then processed by using ML classifiers such as support vector machines (SVM)[14,15], random forests[16], and decision trees[17].

Several research works have introduced video surveillance frameworks to detect suspicious activities, including theft and violence. Techniques such as color histograms, local binary patterns (LBP), and optical

[1]SCSE, Galgotias University, Greater Noida, Uttar Pradesh, India. [2]SCSET, Bennett University (The Times Group), Greater Noida, Uttar Pradesh, India. [3]CSE, Symbiosis Institute of Technology, Pune, Maharashtra, India. [4]CSE, ABV-IIITM, Gwalior, Madhya Pradesh, India. [5]CSED, Mahamaya College of Agricultural Engineering And Technology, Dulla Pur, Uttar Pradesh, India. [6]ECE Department, Manipal University Jaipur, Jaipur, Rajasthan, India. ✉email: deepika.bansal@jaipur.manipal.edu

flow-based descriptors have improved accuracy[18,19]. Additionally, optical flow techniques, including sparse and dense flow algorithms, have been explored to model human motion for behavior recognition[20,21]. Sparse optical flow focuses on selected key points, whereas dense optical flow considers all pixels in a frame, making it more computationally expensive but effective for detecting movement[19,22,23]. In[24], a conventional deep learning model is integrated with the traditional classifier to detect human actions in RGB stream. In[14], a video surveillance system was proposed to identify various risk situations in stores by recognizing suspicious customer behavior. This system used global color histograms, local binary patterns, and histograms of oriented gradients to represent the action and incorporated SVM for action classification. In[15], the authors used an inadequate joint learning structure to detect abnormal events in video transmissions. They extracted features using HOF, HOG, and MBH techniques and used the Support Vector Data Description (SVDD) technique to identify unusual abnormal activities. Although traditional techniques such as color histograms, LBP, and optical flow descriptors have proven to be effective, they require human calibration and face difficulties with complex motion fluctuations.

Deep learning in surveillance

Deep learning methods such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) provide a more flexible solution by autonomously acquiring spatial and temporal features, thereby significantly advancing the field of action recognition. These models improve action recognition accuracy, and reduce the reliance on manually generated feature extraction. Here, the researchers have increasingly focused on end-to-end trainable models that automatically learn spatial and temporal features from video sequences. In[25], a CNN-based framework was proposed to integrate spatial and motion signals for human activity recognition. Other studies have employed two-stage methodologies, where the first stage extracts motion features using optical flow, and the second stage applies machine learning classifiers for activity recognition[26,27].

Recent approaches have combined CNNs with LSTMs to capture both spatial and temporal dependencies. The introduction of dual-stream architectures, where one stream processes RGB frames for spatial features and the other processes optical flow frames for motion features, has improved recognition accuracy[28,29]. InceptionV3, a deep CNN model, has been employed for feature extraction due to its efficiency in handling large-scale visual data. Researchers have also explored 3D convolutional networks and multiple-instance learning (MIL) strategies to improve anomaly detection in videos[29,30]. In[31], a hierarchical 3D convolutional network is proposed to classify abnormal human activities in video sequences. In [32], an activity tracking system called EAGLE is proposed for human activity recognition in occluded environments with improved Generative Adversarial Imputation Network (GAIN) and Bi-LSTM network. Recent trends in human activity recognition emphasizes the significance of attention-based spatio–temporal architectures, including the Bi-LSTM model with attention [33] and the ResDLCNN-GRU attention network [34].

Gap analysis

Despite breakthroughs in surveillance technology, many problems remain that impede the efficacy and dependability of contemporary systems. Current methodologies rely on manually designed features, which often do not generalize well across different environments and activity changes. Machine learning approaches in contemporary research have enhanced accuracy through feature extraction and classification. Nevertheless, these methods still require a lot of feature engineering and manual adjustments, which limits their ability to adjust to changing surveillance conditions[1,30]. Deep learning methods like CNN and RNN, which are based on autonomous learning, have significantly enhanced the detection of human activities. However, their difficulties with varying lighting and real-time processing limitations impact their overall reliability in practical applications. Additionally, a shortage of large, diversified, and well-annotated datasets is another key impediment here. Therefore, we focus on creating effective real-time processing algorithms, adaptive learning strategies, and theft-diverse datasets to overcome these bottlenecks and improve the scalability and accuracy of monitoring systems.

Motivation

The proposed framework addresses such challenges by introducing a hybrid deep learning approach that integrates InceptionV3 for spatial feature extraction and LSTM for temporal modeling. By leveraging both RGB and optical flow streams, the system enhances recognition accuracy for theft activities in indoor surveillance. Additionally, the proposed dataset focuses specifically on theft detection, filling the gap in publicly available datasets for this task. This research contributes to advancing automated surveillance by improving real-time human activity recognition with minimal human intervention.

This study specifically focuses on the recognition of theft activity, a critical aspect of security monitoring. Existing human activity recognition (HAR) methods leverage deep learning architectures such as CNNs, RNNs, and LSTM networks[7,35,36]. However, these approaches often struggle with dynamic environments, occlusions, and variations in human behavior. The proposed research aims to bridge these gaps by leveraging transfer learning and a hybrid deep learning framework to improve recognition performance in surveillance videos.

Contribution

This study introduces a novel, human-intervention-free framework for real-time theft detection, incorporating a dual-stream deep learning approach. The major contributions of this work include:

- **Proposed Approach:** A hybrid deep learning framework that integrates spatial and motion-based feature extraction using RGB and optical flow frames. The approach utilizes a pre-trained InceptionV3 model for extracting visual features and an LSTM network for capturing temporal dependencies in video sequences.

- **Dataset Generation:** A new, specifically curated dataset focused on theft activity detection in indoor surveillance environments. This dataset is carefully collected and structured to enhance the training and evaluation of the proposed HAR system.
- **Comparative Analysis with State-of-the-Art Methods:** Extensive experimental evaluations and comparisons against existing methods are conducted using the proposed dataset and standard benchmarks. The system demonstrates superior accuracy in theft detection.

Paper organization

The paper is organized as follows: "Methodology" presents a comprehensive literature review of previous work on human activity recognition systems. Section 3 details the proposed HAR framework for theft detection, outlining the methodology and system architecture. Section "Experimental results and analysis" discusses the proposed approach's experimental results and performance evaluation on both custom and benchmark datasets. Finally, Section "Conclusion" concludes the paper and highlights potential directions for future research.

Methodology

Framework overview

This paper presents a human activity recognition (HAR) surveillance framework designed for real-time identification of theft and normal behaviors as illustrated in Fig. 1. The HAR framework follows a pipeline of 3 levels: pre-processing, feature extraction, and classification. The pre-processing uses video frames as information and concentrates on groupings/outlines from them. The subsequent step utilizes the feature extraction method to obtain each succession's pertinent highlights. Furthermore, these elements are passed to fabricate a classifier in the characterization unit, which recognizes human activities.

Pipelining level

The work begins with a preprocessing phase, where video recordings are decomposed into frames at a fixed frame rate (30 fps). Each frame is resized to a uniform resolution of 299 × 299 pixels to standardize the input dimensions. Concurrently, optical flow clips are generated from consecutive RGB frames using dense motion estimation to capture temporal movement patterns. These preprocessing steps result in two synchronized input streams: the RGB stream, representing appearance-based features, and the optical stream, capturing motion dynamics.

Feature extraction level

Each stream is processed through a deep InceptionV3 model, which serves as a feature extractor. The model generates spatial features (F1) from the RGB stream and motion features (F2) from the optical stream. The optical frame generation from RGB images using optical flow algorithm is represented in Fig. 2. These features are then fused using average pooling to form a hybrid representation. The resulting combined feature vector is passed to a deep recurrent neural network based on Long Short-Term Memory (LSTM) units. This network models the temporal dependencies across video frames and uses a SoftMax activation function in the final layer to classify the observed activities as either theft or normal.

Algorithm description

The detailed workflow of the proposed HAR method is outlined in **Algorithm** 1. Mathematically, let the input video sequence frames are given by $V = \{f_1, f_2, \ldots, f_n\}$, where each frame $f_i \in \mathbb{R}^{H \times W \times 3}$ is a color image resized to a standard dimension. From this sequence, motion information is extracted by computing optical flow between consecutive frames, denoted as $o_i = \text{OpticalFlow}(f_i, f_{i+1})$, yielding a sequence $O = \{o_1, o_2, \ldots, o_{n-1}\}$. Both RGB and optical frames are passed into two separate streams of the InceptionV3 model to extract spatial and temporal features, respectively: $F1_i = \phi_{\text{RGB}}(f_i)$ and $F2_i = \phi_{\text{Optical}}(o_i)$, where ϕ represents the InceptionV3 feature extraction function. These features are then fused using average pooling

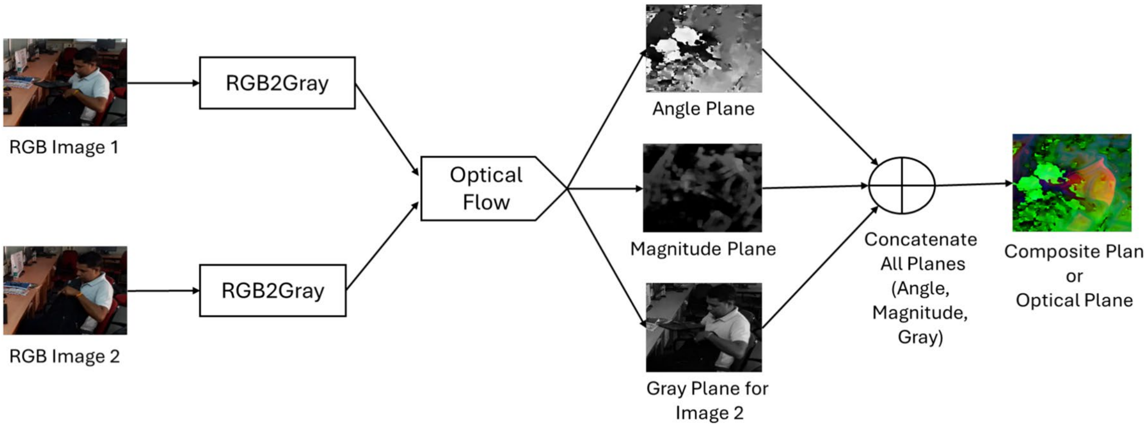


Fig. 1. Optical image generation workflow.

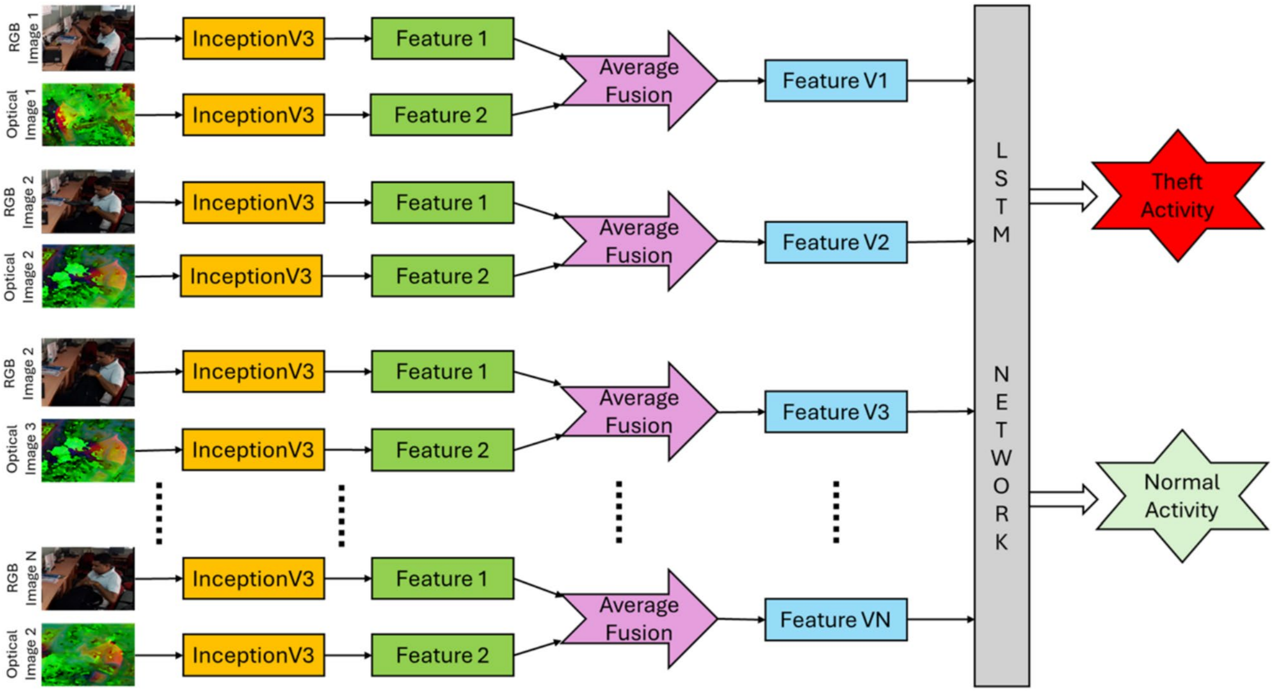


Fig. 2. Proposed method workflow.

to form a hybrid representation $F_i = \frac{F1_i + F2_i}{2}$. The resulting feature sequence $\{F_1, F_2, \ldots, F_{n-1}\}$ is fed into a Long Short-Term Memory (LSTM) network to capture temporal dependencies, producing hidden states $H = \text{LSTM}(F_1, \ldots, F_{n-1})$. The final hidden state h_{final} is passed through a fully connected layer followed by a SoftMax activation function to predict the class probabilities: $\hat{y} = \text{Softmax}(Wh_{\text{final}} + b)$, where $\hat{y} \in \mathbb{R}^2$ corresponds to the predicted likelihood of the input being either a normal or a theft activity. This framework is specifically optimized for identifying anomalous actions, such as theft, in surveillance environments like laboratories.

```
Input: Video $V = \{f_1, f_2, ..., f_n\}$ (sequence of frames)
Output: Activity label $\in$ {Normal, Theft}
Load pre-trained InceptionV3 model
Initialize LSTM network
for each frame $f_i$ in $V$ do
    Resize $f_i$ to $299 \times 299$
end for
Generate optical flow images $\{o_1, ..., o_{n-1}\}$
for each frame $f_i$ in RGB stream do
    $F1_i \leftarrow$ InceptionV3.extract_features($f_i$)
end for
for each flow frame $o_i$ do
    $F2_i \leftarrow$ InceptionV3.extract_features($o_i$)
end for
for $i = 1$ to $n - 1$ do
    $F_i \leftarrow \text{Average}(F1_i, F2_i)$
end for
$O \leftarrow \text{LSTM.forward}(\{F_1, ..., F_{n-1}\})$
Label $\leftarrow \text{Softmax}(O)$
return Label
```

Algorithm 1. Dual-Stream Deep Learning for Human Activity Recognition

The core technical modules used in the proposed algorithm are detailed as follows.

Optical image generation

Optical flow [13,37] is significantly utilized in many applications while working with movement streams of articles in the recording of videos. The optical flow developments include vectors from an arrangement of the back-to-back casing to depict movement data for a moving item. According to Gunnar Farneback's algorithm [17], the optical flow between two back-to-back images is assessed as follows:

Consider a period example *t*; let a point on this edge be addressed with the position coordinates *x* and *y*. At any next time, or occasion, a similar point in the following block takes an alternate situation with an extremely more modest relocation δx and δy in the X and Y bearings, respectively. Accordingly, the refreshed place of this pixel in the following casing is $I(x + \delta x, y + \delta y, t + \delta t)$, expecting that the force of the point *I* continue to be of similar worth during the development in back-to-back steps. Accepting little movement between outlines, the Taylor series development is:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \tag{1}$$

where

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \approx 0 \tag{2}$$

According to the steady brilliance presumption. By partitioning δt, the optical stream condition is:

$$I_x \cdot u_x + I_y \cdot v_y = 0 \tag{3}$$

where $u_x = \frac{\delta x}{\delta t}$ and $v_y = \frac{\delta y}{\delta t}$ are speed vectors. Likewise, $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, and $\frac{\partial I}{\partial t}$ are the frame slopes alongside the level pivot, vertical hub, and time, respectively. The above condition addresses the optical stream imperative for two questions (i.e., assuming a frame joins *N* pixels, it gives *N* conditions to 2*N* obscure).

Feature extraction (InceptionV3)

The Inception [1,23] model is a CNN-based architecture mainly used for feature extraction and classification. It aims to act as a staggered highlight extractor and reduce computational efficiency. Figure 3 shows the deep architecture of the InceptionV3 module used to extract features from images. To reduce computing costs without affecting the ability of representation, a factorization strategy is used in which larger convolutions (e.g. 5 × 5) are decomposed into smaller convolutions (two 3 × 3). InceptionV3 also includes asymmetric convolutions (for example, 1 × 7 followed by 7 × 1) to capture elongated features and maintain efficiency. The architecture uses auxiliary classifiers during training to eliminate gradients and improve convergence. Stack normalization is widely used to accelerate training and improve generalization. With over 24 million parameters and 48 deep layers, InceptionV3 achieves top 1 accuracy of 78.1% and top 5 accuracy of 94.2% with image-protected data records. Its modular design allows for multi-scale functional extraction, making it particularly suitable for capturing spatial hierarchies in complex visual scenes, such as those found in tasks to recognize human activity.

Motion and appearance fusion

Another crucial concept in the network is matrix size reduction, which involves downscaling feature maps using convolution with a stride value and a max pooling operation. The network becomes more efficient and cost-effective as a result of this link-based matrix size reduction procedure.

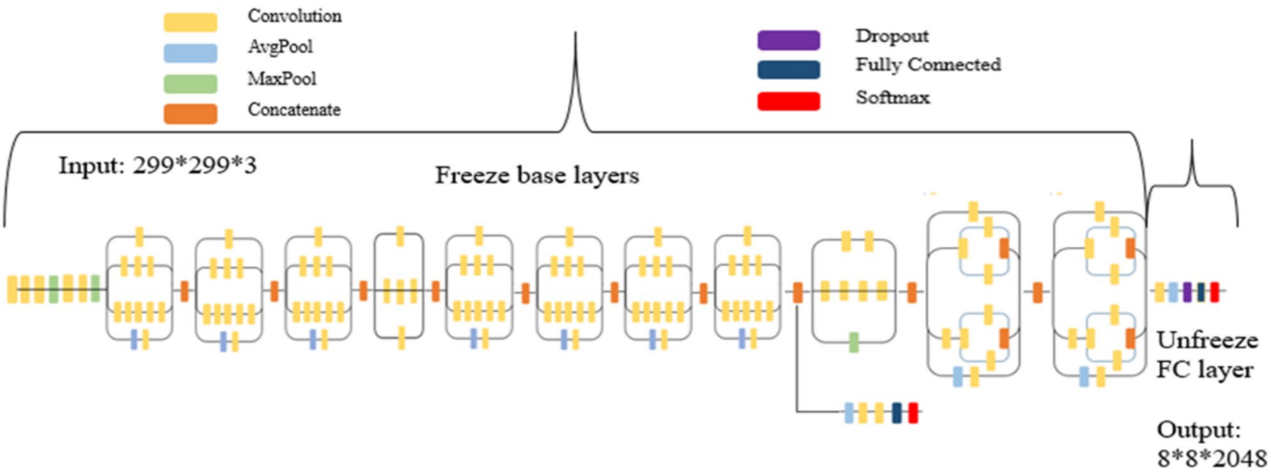


Fig. 3. Transfer learning based on InceptionV3 architecture.

Temporal modeling (LSTM)
A recurrent neural network (RNN) [7,8,38] is a type of neural network architecture that incorporates a form of memory known as transient memory. However, RNNs are limited to long-haul conditions because of their insufficient ability to convey data from before to afterwards. To overcome this, a variation of the RNN-based model LSTM [23,28,39,40] is used to learn longer input videos more effectively. It is equipped for learning long-haul conditions and can handle single images and numerous information arrangements (video) effectively.

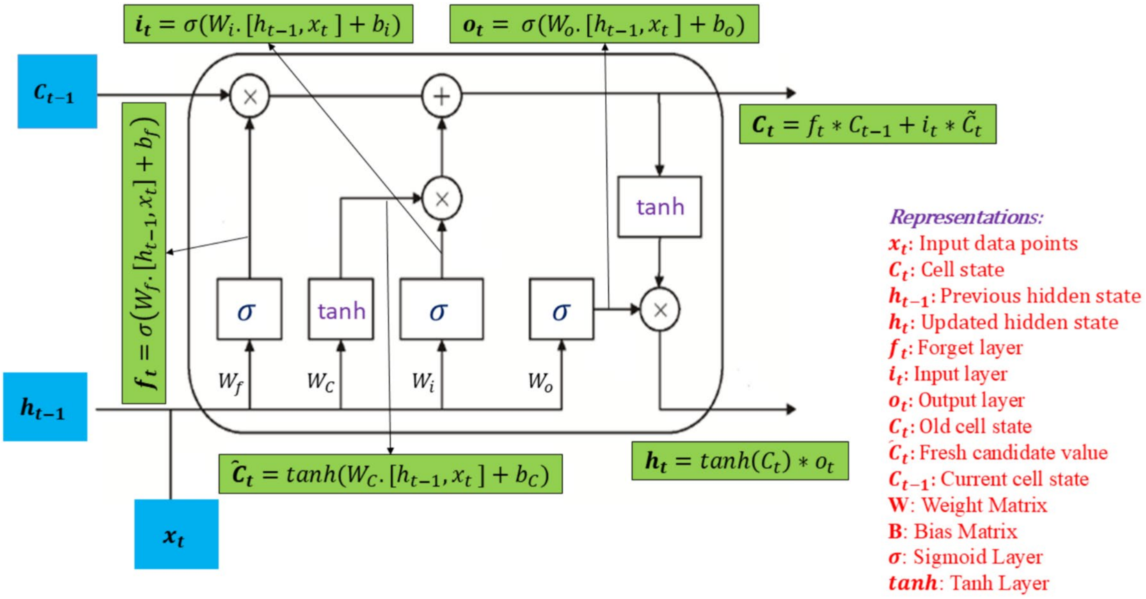
Because of their unique ability to perceive and represent sequential patterns in data, the LSTM networks perform well as classifiers. Each input instance is defined as a series of features or tokens for classification, as shown in Fig. 4. Each frame or sequence of frames is handled as a time step in the LSTM in the context of video classification. The memory cell and gating processes of the LSTM are critical in keeping important information while eliminating irrelevant details, allowing the model to identify meaningful patterns over time. LSTMs can effectively identify complicated temporal correlations, such as object movements, activities, and scene changes, by studying the sequential evolution of video frames, which is critical for accurate video classification. The LSTM's final hidden state can be sent through a fully connected layer and a softmax layer to provide classification predictions based on the learned features.

Dataset description

The need for dataset creation arises from the challenge of recognizing abnormal behavior in universities and colleges because the authorities are facing the issue of theft in labs, which cannot be monitored manually. Moreover, our literature survey shows a standard lab lifting dataset is absent. Therefore, the proposed dataset is required to cover the abnormal events present in the lab, which include actions such as stealing items, namely keyboards and computer mice from the lab. Universities and educational institutions can use this dataset to monitor laboratory theft-related activities.

A dataset was generated primarily for general scenarios that occur in college or institute lectures. Sometimes, students may steal computer equipment, such as mice and keyboards, in computer laboratories. However, manually identifying such students is difficult, resulting in financial losses for the universities. To help limit such unusual events and their accompanying losses, we created a dataset called "Theft Activities in Laboratories: Lab Lifting". This dataset was created by capturing video recordings at a frame rate of 30 frames per second with a 13-megapixel camera with a resolution of 640×480 . Recordings were made from different camera angles and under slightly varying lighting conditions to approximate real-world settings. The events in the lab-lifting dataset were divided into two types: normal actions and theft actions. Figure 5 represents some samples of videos recorded from the laboratorylift dataset.

This dataset's normal class includes students' activities such as working on computers, using mobile phones, resting in front of a computer, and instances of empty classrooms. At the same time, anomalous classes involve unusual activities like putting a mouse under clothing or in a bag and putting the keyboard in a bag. The distribution of the theft actions in the laboratory dataset is shown in Table 1. The dataset consists of 178 video clips, 10 seconds long, efficiently representing a person's stealing and usual actions in the computer laboratory. To ensure variability and enhance generalization, the dataset includes recordings from multiple individuals under different camera angles and slightly varying lighting conditions, simulating real-world diversity in lab environments. In this dataset, 72 clips represent normal/usual event occurrences, while 86 clips represent theft events in the laboratory. Out of 178 video clips, 133 clips were used for training, whereas 45 clips were utilized for testing.



$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$
$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$
C_{t-1}
$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$
$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
tanh
σ
tanh
σ
σ
W_f
W_C
W_i
W_o
h_{t-1}
$\hat{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C)$
$h_t = tanh(C_t) * o_t$
x_t
Representations:
x_t: Input data points
C_t: Cell state
h_{t-1}: Previous hidden state
h_t: Updated hidden state
f_t: Forget layer
i_t: Input layer
o_t: Output layer
C_t: Old cell state
$\hat{C}_t$: Fresh candidate value
C_{t-1}: Current cell state
W: Weight Matrix
B: Bias Matrix
σ: Sigmoid Layer
$tanh$: Tanh Layer

Fig. 4. The architecture of the LSTM Network.

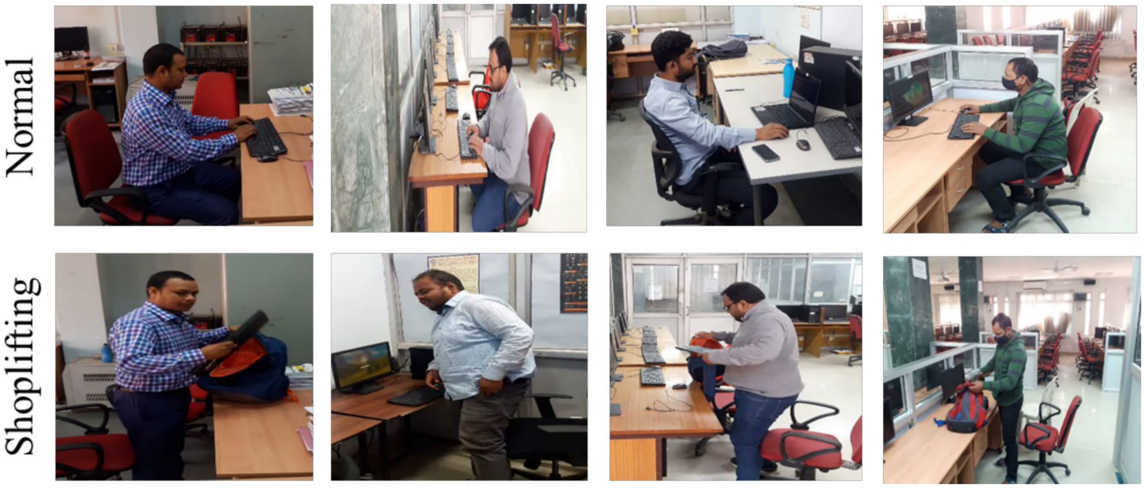


Fig. 5. Instances of synthesized dataset.

Class	Mode	No. of clips	No. of videos
Lab normal videos	Testing	24	92
	Train	68	
Lab theft videos	Test	21	86
	Train	65	

Table 1. Distribution of dataset across different modes.

Experimental results and analysis

Experimental setup

All experiments were carried out on a system running Ubuntu 16.0 LTS and powered by an Intel Core i7-4005 CPU. The machine came with 16 GB of RAM and a 4 GB NVIDIA GeForce graphics card (1650 GTX). CUDA version 10.0.130 was used to optimize the utilization of the graphics processing unit (GPU) for faster processing. The experiments were conducted using Python 3.6 and the TensorFlow-GPU 1.14 and Keras 2.2 packages. These tools were intended to simplify the proposed framework's training and simulation. The model was trained using the 'Adam' optimizer with a learning rate of 0.0001, which enabled stable and efficient convergence of the network parameters. Section 4.1 goes through the specifics of the training and testing data used in the studies.

In the context of human behavior classification, accuracy and loss are fundamental evaluation metrics that provide insights into the performance of the classification model. Here is a description of each metric, along with its formula:

Accuracy: The proportion of correctly categorized instances in the dataset is measured based on the accuracy as stated in Equation 4. Accuracy describes how well a model can predict the many types of human behaviors, such as normal activities and pathological activities, in the context of classifying human behavior.

$$\text{Accuracy} = \frac{\text{Number of corrected behavior predictions}}{\text{Total number of behavior predictions}} \tag{4}$$

Loss: Loss, often referred to as the "cost" or "error," quantifies how well the model's predictions match the actual labels. It represents the discrepancy between the predicted values and the ground truth labels. In the context of human behavior classification, minimizing the loss helps the model learn to make better predictions over time. Common loss functions include mean squared error (MSE) for regression tasks and cross-entropy loss (CEL) for classification tasks. For human behavior classification, cross-entropy loss is frequently used as it is well-suited for multiclass classification problems, as represented in Equation 5.

$$\text{CEL} = \frac{-1}{M}\sum_{j=1}^{M}(y_j\log(p_j) + (1-y_j)\log(1-p_j)) \tag{5}$$

where M is the number of samples, y_j is the truth label, and p_j is the probability using the sigmoid function. These metrics, accuracy and loss, quantitatively measure the model's performance in human behavior classification. By analyzing both accuracy and loss, researchers and practitioners can assess the model's ability to distinguish between different human behavior categories and make informed decisions about model improvements or adjustments.

BS	Epochs	TA	TL	VA	VL	TT
4	95	99.99	0.0003	87.79	1.12	22.71
8	82	100	0.0004	85.36	1.08	20.10
16	124	100	0.0004	87.80	0.752	18.20
32	97	100	0.0013	90.24	0.686	7.20

Table 2. Experimental outcomes for S.L. of 300 with no augmentation.

BS	Epochs	TA	TL	VA	VL	TT
4	89	93.43	0.1780	85.71	0.454	44.02
8	93	99.61	0.0001	91.86	0.3151	39.02
16	80	100	0.0003	90.24	0.499	35.20
32	90	99.22	0.0120	89.28	0.593	14.20

Table 3. Experimental outcomes for S.L. of 300 with augmented data.

BS	Epochs	TA	TL	VA	VL	TT
4	87	97.87	0.067	88.37	0.8059	13.41
8	92	99.22	0.001	88.37	0.6610	10.75
16	125	100	0.013	86.04	0.7498	8.94
32	131	96.89	0.134	81.39	0.4155	5.32

Table 4. Experimental outcomes for S.L. of 150 with no augmentation.

Performance evaluation

The proposed model was fine-tuned by analysing its predictive modelling behaviour utilising diverse examples. The sequence length represented the number of frames used to address an activity. The theft data in this scenario was made up of 10 seconds long video clips collected at 30 frames per second, yielding a sequence length value of 300 for the initial trials. On the other hand, processing 300 frames for each clip in the vast collection of theft videos was computationally intensive. As a result, the sequence length was reduced by 50% in the following phase of the trials. Another instance was used to double the number of clips in the video sequence dataset by changing the spatial orientations horizontally. This suggested organization works for divergent group measures and addresses the unique conduct conditions of the proposed model.

In the initial case, the suggested model was applied with a sequence length (S.L.) of 300 without any augmentation. The time required for feature extraction was around 39 min and 15 s. Table 2 displays the experimental results of the suggested architecture. Here, TA, TL, VA, VL and TT are training accuracy, training loss, validation accuracy, validation loss and training time (in seconds/epoch). In the evaluation, it was observed that the suggested model outperformed other models on a 32-batch size, with a training accuracy as high as 100% and validation accuracy as high as 90.24%. Furthermore, the model displayed fewer training and validation losses, alluding to robust network behavior.

The second scenario used longer video sequences, resulting in sequences 300 frames longer. The feature analysis procedure required much more time when using the suggested model; it took approximately 1 hour, 47 minutes, and 29 seconds. Table 3 shows the second scenario's outcomes and the suggested models' effectiveness. The model achieved up to 99.61% training accuracy and 91.86% validation accuracy on an eight-node cluster. The model exhibited steady behavior, as shown by the reduced training loss and reasonable validation loss, comparable to the first case.

Longer sequence times in the initial and subsequent tests required more computational resources and memory. In this instance, the sequence length was divided into segments within the video clips. The third scenario used less memory and computational power than the first and second scenarios. The duration of the 150-frame sequences in the third scenario was identical to that in the model that was being proposed. As a result, the feature extraction process took about 27 minutes and 13 seconds less time. Table 4 presents the trial data for the third case and displays how well the proposed model performed. On an eight-batch size, the model had the highest accuracy, with a training accuracy of 99.22% and a validation accuracy of 88.37%.

For the suggested model, the final scenario used longer video clips with a sequence duration of 150 frames. Table 5 displays the experiment's outcomes for this case, along with the confusion matrix's parameters. In this scenario, feature extraction took about 1 hour, 2 minutes, and 56 seconds. In Comparison to previous cases, the suggested model achieved high accuracy, with 91.86% accuracy in training and 91.86% accuracy in validation with a batch size of eight. On the other hand, the model provided the highest false positive and negative instances for a batch size of four, indicating that its performance is outstanding for a batch size of eight. The framework proposed for the final scenario behaves consistently with the first three cases.

BS	Epochs	TA	TL	VA	VL	TT
4	125	93.75	0.160	70.93	0.690	76.84
8	120	93.75	0.219	91.86	0.220	42.16
16	120	91.40	0.241	86.27	0.367	23.88
32	224	95.70	0.154	87.20	0.388	12.45

Table 5. Experimental outcomes for S.L. of 150 with augmented data.

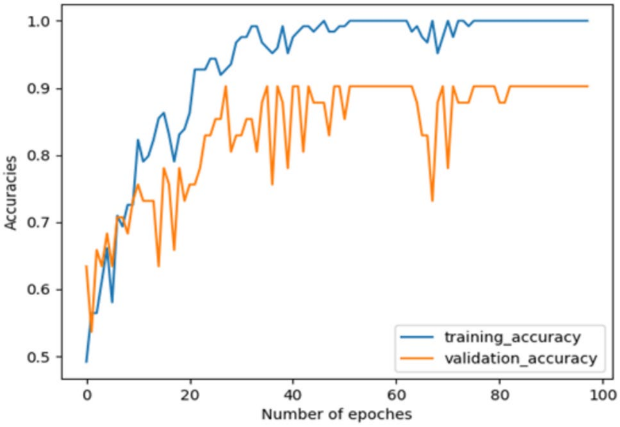


Fig. 6. Accuracy trade-off for S.L. of 300, without augmentation, and batch size of 32.

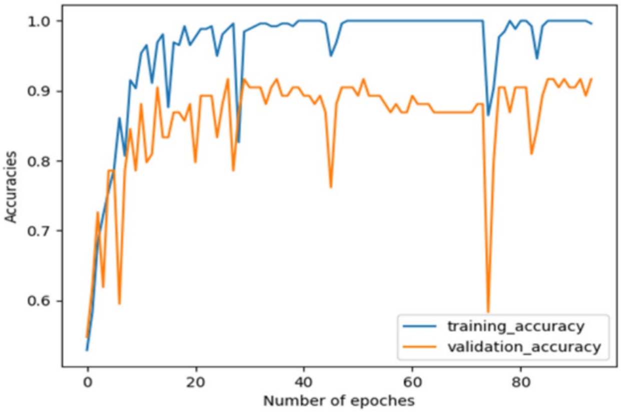


Fig. 7. Accuracy trade-off for S.L. of 300, with augmentation, and batch size of 8.

Comparative evaluation

The proposed model, which is trained on a subset of the combined theft dataset with enhanced video clip duration, obtained the highest validation accuracy among other models when employing a batch size of 32 after analysing the experimental outcomes. Figures 6, 7, 8 and 9 show the trade-off between accuracy for the same model. The outcomes show that the model had improved accuracy and reduced loss, demonstrating the network's favourable behaviour throughout the training period.

The suggested model, trained on a subset with a sequence length of above 50% by expansion, surpassed other models in terms of performance, which can be inferred from the experiment's findings of the current shop theft data. It is clear from the graphs that the proposed model achieved high accuracy during training and modest accuracy during validation. The model additionally showed reduced loss during training and slight loss during validation.

Comparison with existing methods

Finally, the existing state-of-the-art methods that have been extensively used in human activity and anomaly detections are compared with our approach, as presented in Table 6. We achieved a higher accuracy of 91.86% compared to Method [1], Method [38], Method [17], Method [28] and Method [32]. The experimental results demonstrate

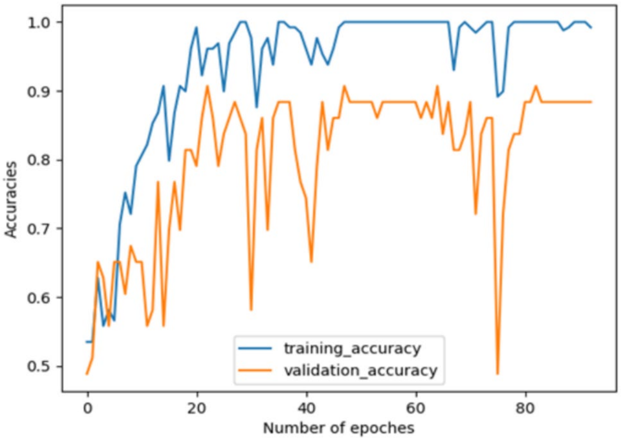


Fig. 8. Accuracy trade-off for S.L. of 150, without augmentation, and batch size of 8.

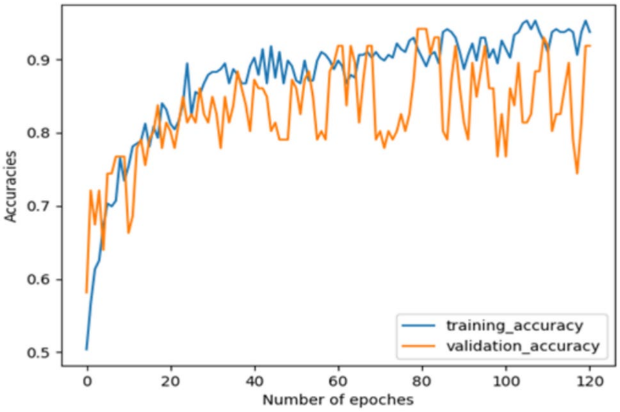


Fig. 9. Accuracy trade-off for S.L. of 150, with augmentation, and batch size of 8.

Method	M[1]	M[38]	M[17]	M[28]	M[32]	PM
Accuracy (%)	89.36	87.23	84.04	88.59	89.36	91.86

Table 6. Comparison of the proposed method with existing methods.

that our proposed approach outperforms existing methods in detecting the theft of retail products through video surveillance. The findings confirm that our dual-stream InceptionV3-LSTM model recorded the highest validation accuracy of 91.86% compared to 3D CNN (89.36%), CNN-LSTM (87.23%), ConvLSTM (88.59%), and EAGLE (89.36%). Although 3D CNN and ConvLSTM are powerful in modeling spatio-temporal correlations, they are computationally expensive and sensitive to long sequence invariance. CNN-LSTM and EAGLE methods, although robust in sequential modeling and occluded views, are second only to the CNN-LSTM in capturing fine-grained motion information. By utilizing complementary RGB-based spatial and optical-flow-driven temporal features combined via LSTM, the developed model exhibits superior detection performance, validating its efficacy for real-world surveillance applications such as theft detection.

Resulting samples and discussion

The experimental results demonstrate the effectiveness of the proposed framework in recognizing human abnormal activities, particularly in theft detection scenarios. The high validation accuracy of 91.86% suggests that the model is capable of learning complex spatial and temporal patterns in human behavior. Additionally, integrating spatial feature extraction via InceptionV3 and temporal pattern learning using LSTM contributed to the model's robustness. The resulting instances after evaluating the proposed model on some unseen videos are shown in the Fig. 10.

One of the framework's strengths is its ability to maintain high accuracy across different batch sizes and sequence lengths. The results in Tables 2, 3, 4 and 5 indicate that the model performs optimally when using an

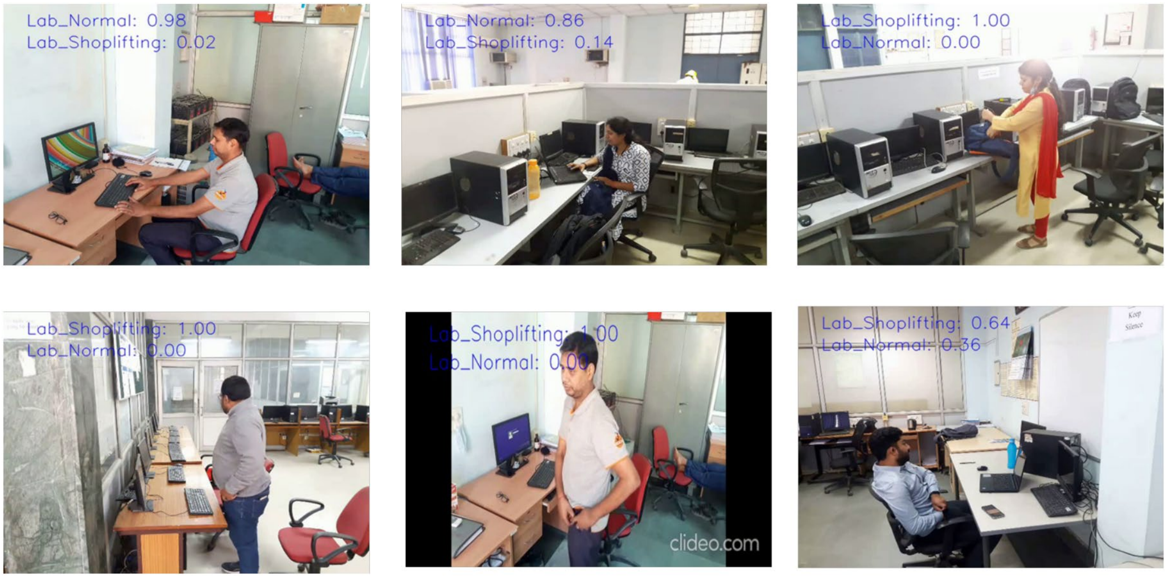


Fig. 10. Resulting samples on unseen videos.

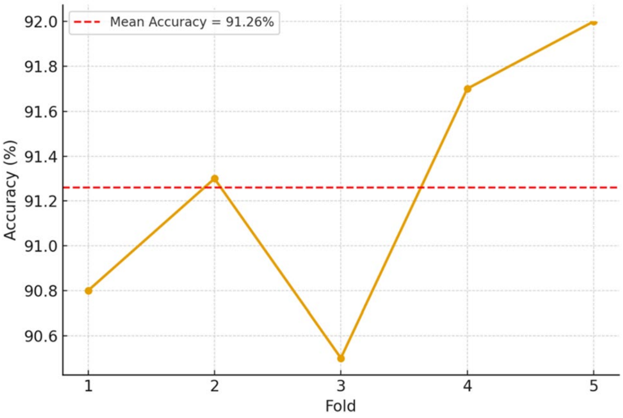


Fig. 11. Accuracy trends across five folds of cross-validation.

appropriate batch size and sequence length. The augmentation techniques applied in the experiments helped enhance the model's generalization capability, reducing the risk of overfitting.

Figure 11 illustrates the accuracy achieved in the five-fold cross-validation for the proposed framework that contains 300 sequence lengths with augmented data. The results demonstrate consistent performance, with precision values ranging from 90.5% to 92.0%. The mean precision of 91.26% (indicated by the dashed red line) highlights the model's stability, and the small variation between folds reflects its strong generalization capability. This statistical validation reinforces the reliability of the experimental findings and supports the effectiveness of the proposed approach.

Despite its strong performance, the proposed framework has certain limitations that warrant further consideration. First, the experiments were conducted on a custom dataset collected in a controlled environment, which may not fully reflect the complexity of real-world surveillance scenarios involving varying lighting conditions, occlusions, background clutter, or diverse camera perspectives.Additionally, the computational cost of processing long video sequences remains challenging, particularly when dealing with high-resolution surveillance footage in real-time applications. Reducing inference time while maintaining accuracy would be a key area for improvement.

Another potential issue is the model's sensitivity to edge cases, such as rapid and unpredictable movements that may not fit typical theft patterns. In such scenarios, false positives or false negatives could increase, impacting the model's reliability in practical deployments. Future research should explore incorporating additional modalities, such as depth information or multi-camera perspectives, to enhance the framework's robustness.

Overall, while the proposed framework exhibits strong performance, further improvements in data diversity, computational efficiency, and robustness to edge cases will be necessary for real-world adoption.

Ablation study

The proposed dual-stream theft recognition framework involves three main computational stages. The first stage, optical flow generation, has a complexity of $O(N \cdot H \cdot W)$ and is efficient when frames are resized to 299×299, allowing processing on standard GPUs. The second stage, feature extraction using InceptionV3, has complexity $O(N \cdot H \cdot W \cdot C)$, where processing a 300-frame clip took about 2,355 seconds (approximately 39 minutes) without augmentation and around 6,449 seconds (about 1 hour 47 minutes) with augmentation. The third stage, temporal modeling with LSTM, has complexity $O(N \cdot d^2)$. This step is relatively important because it scales linearly with sequence length.

In practical terms, for 300-frame sequences without augmentation and a batch size of 32, the per-epoch training time was 7.20 seconds with a validation accuracy of 90.24%. With augmentation and a batch size of 8, training time rose to 39.02 seconds, improving accuracy to 91.86%. For shorter sequences of 150 frames, training without augmentation (batch size 8) took 10.75 seconds per epoch with 88.37% accuracy, while with augmentation it required 42.16 seconds per epoch, achieving 91.86% accuracy. These findings confirm that the proposed approach delivers competitive accuracy while maintaining efficient training times.

The further ablation analysis summarized in Table 7 reveals each core module's individual and cumulative impact within the proposed HAR framework. When using only the RGB stream, the model achieves an accuracy of 89.66%, indicating that static appearance-based features extracted by InceptionV3 are effective to some extent in identifying human actions. However, replacing the RGB stream with optical flow alone results in a slightly lower accuracy of 88.51%, demonstrating that motion information alone lacks sufficient spatial context. This drop in accuracy highlights a limitation of using motion-only (optical flow) features, as they lack spatial detail and object-specific visual cues that are critical for discriminating against subtle actions such as object theft. Importantly, the full configuration of the proposed model, integrating both RGB and optical flow streams, feature fusion, and LSTM for temporal modelling, achieves the highest accuracy of 91.86%. This confirms the complementary nature of spatial and temporal features and highlights the importance of motion-appearance fusion and sequence modeling for robust activity recognition. The role of the LSTM is particularly significant, as it captures temporal dependencies across the fused features, enhancing the system's ability to distinguish complex actions like theft from normal activity.

Complementary to component-level evaluation, Fig. 12 examines the effect of different optimization algorithms on model convergence and performance. The Adam optimizer delivers the best result, achieving 91.86% accuracy, underscoring its effectiveness in adapting learning rates for each parameter and handling sparse gradients, a common characteristic in video-based models. AdamW, a variant that decouples weight decay from the gradient update, follows closely with 90.47%, while traditional methods such as SGD and AdaGrad perform less favorably at 86.74% and 87.05%, respectively. This can be attributed to their sensitivity to hyperparameter tuning and slower convergence. Overall, the optimizer choice significantly influences model performance, especially in deep sequential architectures, where adaptive learning strategies better accommodate the temporal complexities of HAR tasks.

Conclusion

Summary of contributions

This research introduces a comprehensive framework for recognizing human abnormal activities, incorporating spatial and temporal variations. The framework leverages deep learning techniques to refine and optimize feature extraction, utilizing a combination of CNN and RNN models. Specifically, InceptionV3 is employed for spatial feature extraction from sequential frames, while LSTM is used to classify activities, distinguishing between theft-related and normal behaviors. A novel approach involving composite frames for motion feature extraction is introduced, integrating RGB and optical flow planes to enhance activity recognition. The study focuses on theft activities captured in laboratory environments, providing a specialized dataset for training and evaluation.

Key findings

Experimental results demonstrate that the proposed framework effectively differentiates between various human activities, achieving an accuracy of 91.86%, outperforming existing state-of-the-art methods. The model exhibits robust performance across different testing scenarios, with enhanced feature extraction contributing to improved classification accuracy. The Integrating spatial and temporal information allows the system to recognize complex activity patterns reliably.

Future work

Future research directions include expanding the dataset to cover a broader range of indoor and outdoor human activities to improve generalization. Further, we can highlight the possibility of integrating unattended learning techniques such as latent spatial modelling networks and action translation networks to expand the framework of more broadly autonomous anomaly recognition applications. Additionally, enhancing the accuracy and

Configuration	RGB Stream	Optical Flow	InceptionV3	Fusion	LSTM	Accuracy (%)	Precision (%)	Recall (%)
Baseline (RGB Only)	✓	–	✓	–	✓	89.66	88.41	90.89
Optical Flow Only	–	✓	✓	–	✓	88.51	87.10	89.42
Full Model (RGB + Optical + Fusion)	✓	✓	✓	✓	✓	**91.86**	**91.15**	**91.88**

Table 7. Evaluating the impact of different modules in the proposed HAR framework.

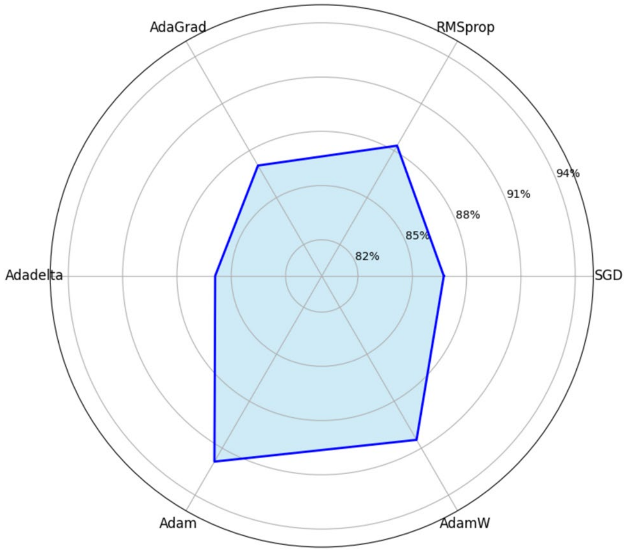


Fig. 12. Performance comparison of the proposed HAR model for different optimizers.

efficiency of the model through advanced deep learning architectures and optimized feature fusion techniques is a key focus. Another promising avenue is deploying the framework in real-world surveillance systems, enabling real-time activity recognition in diverse environments. Furthermore, incorporating semantic scene understanding and multi-modal fusion strategies could enhance the model's capability to interpret human behaviors more effectively.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 7 August 2025; Accepted: 31 October 2025
Published online: 25 November 2025

References

1. Ansari, M.A., & Singh, D.K. An expert eye for identifying shoplifters in mega stores. In *Proceedings of the 4th International Conference on Innovative Computing and Communication (ICICC 2021)*, New Delhi, India (2021).
2. Varshney, N. Deep learning in human activity recognition from videos: A survey. In *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT*, pp. 335–346. Springer, Singapore (2022).
3. Irfanullah, H. T., Iqbal, A., Yang, B. & Hussain, A. Real-time violence detection in surveillance videos using convolutional neural networks. *Multimedia Tools Appl.* **81**, 38151–38173 (2022).
4. Pareek, P. & Thakkar, A. Rgb-d based human action recognition using evolutionary self-adaptive extreme learning machine with knowledge-based control parameters. *J. Ambient. Intell. Humaniz. Comput.* **14**, 939–957 (2021).
5. Hussain, T. et al. Improving source location privacy in social internet of things using a hybrid phantom routing technique. *Comput. Secur.* **123**, 102917 (2022).
6. Singh, R., Kushwaha, A. K. S. & Chandni, S. R. Recent trends in human activity recognition–a comparative study. *Cogn. Syst. Res.* **77**, 30–44 (2023).
7. Pienaar, S.W., & Malekian, R.: Human activity recognition using lstm-rnn deep neural network architecture. In *Proceedings of the Wireless Africa Conference (WAC), Pretoria, South Africa*, pp. 1–5 (2019).
8. Xia, K., Huang, J. & Wang, H. Lstm-cnn architecture for human activity recognition. *IEEE Access* **8**, 56855–56866 (2020).
9. Varshney, N. Combining electrocardiogram signal with accelerometer signals for human activity recognition using convolution neural network. *J. Phys: Conf. Ser.* **1947**, 012037 (2021).
10. Singh, D. K., Paroothi, S., Rusia, M. K. & Ansari, M. A. Human crowd detection for city wide surveillance. *Proc. Comput. Sci.* **171**, 350–359 (2020).
11. Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **103**, 60–79 (2013).
12. Singh, D. & Mohan, C. K. Graph formulation of video activities for abnormal activity recognition. *Pattern Recogn.* **65**, 265–272 (2017).
13. Ladjailia, A., Bouchrika, I., Merouani, H. F., Harrati, N. & Mahfouf, Z. Human activity recognition via optical flow: Decomposing activities into basic actions. *Neural Comput. Appl.* **32**, 16387–16400 (2020).
14. Arroyo, R., Yebes, J. J., Bergasa, L. M., Daza, I. G. & Almazán, J. Expert video-surveillance system for real-time detection of suspicious behaviours in shopping malls. *Expert Syst. Appl.* **42**, 7991–8005 (2015).
15. Nguyen, T.N., & Ly, N.Q. Abnormal activity detection based on dense spatial-temporal features and improved one-class learning. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*, Nha Trang City, Vietnam (2017).
16. Kumar, K. P. & Sanal, B. R. Human activity recognition in egocentric video using hog, gist and colour features. *Multimedia Tools Appl.* **79**, 3543–3559 (2020).
17. al., G.M.-M.G. Criminal intention detection at early stages of shoplifting cases by using 3d convolutional neural networks. *Computation* **9**(2), 24 (2021).

18. Lingaswamy, S. & Kumar, D. An efficient moving object detection and tracking system based on fractional derivative. *Multimedia Tools Appl.* **79**, 8519–8537 (2018).
19. Aggarwal, J. K. & Ryoo, M. S. Human activity analysis: A review. *ACM Comput. Surv.* **43**(3), 1–43 (2011).
20. Vijeikis, R., Raudonis, V. & Dervinis, G. Efficient violence detection in surveillance. *Sensors* **22**, 2216 (2022).
21. Zhang, M., Hu, H., Li, Z. & Chen, J. Action detection with two-stream enhanced detector. *Vis. Comput.* **39**, 1193–1204 (2022).
22. Tarnec, L. L., Destrempes, F., Cloutier, G. & Garcia, D. A proof of convergence of the horn-schunck optical flow algorithm in arbitrary dimension. *SIAM J. Imag. Sci.* **7**(1), 277–293 (2014).
23. Ansari, M. A. & Singh, D. K. Esar, an expert shoplifting activity recognition system. *Cyber. Inf. Technol.* **22**(1), 190–200 (2022).
24. Basly, H., Ouarda, W., Sayadi, F.E., Ouni, B., & Alimi, A.M. Cnn-svm learning approach based human activity recognition. In *International Conference on Image and Signal Processing*, pp. 271–281 (2020).
25. Feichtenhofer, C., Pinz, A., & 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, NV, USA (2016).
26. Ibrahim, N., Mustafa, M.M., Mokri, S.S., Siong, L.Y., & Hussain, A. Detection of snatch theft based on temporal differences in motion flow field orientation histograms. *Int. J. Adv. Comput. Technol.* **4** (2012).
27. Rashwan, H. A., Garcia, M. A., Abdulwahab, S. & Puig, D. Action representation and recognition through temporal co-occurrence of flow fields and convolutional neural networks. *Multimedia Tools Appl.* **79**, 34141–34158 (2020).
28. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA* (2015).
29. Sultani, W., Chen, C., & Shah, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA* (2018).
30. Ji, S., Xu, W., Yang, M. & Yu, K. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2012).
31. Ansari, M.A., & Singh, D.K. Deep-3dconvnet: A network to detect abnormal activities at megastores. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1–5 (2022). IEEE
32. Kurchaniya, D., & Kumar, S. Eagle: Enhanced activity recognition in occluded environments with improved gain and bi-lstm. *IEEE Trans. Comput. Soc. Syst.* (2025).
33. Kumar, M., Patel, A.K., Biswas, M., & Shitharth, S. Attention-based bidirectional-long short-term memory for abnormal human activity detection. *Sci. Rep.* **13**(1). https://doi.org/10.1038/s41598-023-41231-0 (2023).
34. Dey, A., Biswas, S., & Abualigah, L. Efficient violence recognition in video streams using resdlcnn-gru attention network. *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)* **18**(3), 329–341 (2024).
35. Jayaswal, R., & Dixit, M. A face mask detection system: An approach to fight with covid-19 scenario. *Concurr. Comput. Pract. Exp.* **34**(e7394) (2022).
36. Goel, D., & Pradhan, R.: A comparative study of various human activity recognition approaches. In *IOP Conference Series: Materials Science and Engineering*, vol. 1131 (2021).
37. Kushwaha, A., Khare, A. & Khare, M. Human activity recognition algorithm in video sequences based on integration of magnitude and orientation information of optical flow. *Int. J. Image Graph.* **22**(1), 2250009 (2021).
38. Jayaswal, R. & Dixit, M. A framework for anomaly classification using deep transfer learning approach. *Revue d'Intelligence Artificielle* **35**(3), 255–263 (2021).
39. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden* (2003).
40. Tripathi, R. K., Jalal, A. S. & Agrawal, S. C. Suspicious human activity recognition: A review. *Artif. Intell. Rev.* **50**, 283–339 (2017).

Author contributions

M.A.A. and A.M. wrote the main manuscript text. A.K. and R.J. prepared Figs. 1, 2, 3 and contributed to data analysis. A.S.Y. and L.K. were involved in data collection and methodology design. D.B. supervised the project and performed final manuscript editing. All authors reviewed and approved the final manuscript.

Funding

Open access funding provided by Manipal University Jaipur.

Declarations

Competing interests

The authors declare no competing interests.

Ethical consent

Informed consent was obtained from all the participants involved in the study.

Additional information

Correspondence and requests for materials should be addressed to D.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.