Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.033493
Article

Tech Science Press

Enhanced Deep Learning for Detecting Suspicious Fall Event in Video Data

Madhuri Agrawal* **and Shikha Agrawal**

Department of CSE, University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India
*Corresponding Author: Madhuri Agrawal. Email: madhuriagrawal2000@gmail.com
Received: 18 June 2022; Accepted: 07 November 2022

Abstract: Suspicious fall events are particularly significant hazards for the safety of patients and elders. Recently, suspicious fall event detection has become a robust research case in real-time monitoring. This paper aims to detect suspicious fall events during video monitoring of multiple people in different moving backgrounds in an indoor environment; it is further proposed to use a deep learning method known as Long Short Term Memory (LSTM) by introducing visual attention-guided mechanism along with a bi-directional LSTM model. This method contributes essential information on the temporal and spatial locations of 'suspicious fall' events in learning the video frame in both forward and backward directions. The effective "You only look once V4" (YOLO V4)–a real-time people detection system illustrates the detection of people in videos, followed by a tracking module to get their trajectories. Convolutional Neural Network (CNN) features are extracted for each person tracked through bounding boxes. Subsequently, a visual attention-guided Bi-directional LSTM model is proposed for the final suspicious fall event detection. The proposed method is demonstrated using two different datasets to illustrate the efficiency. The proposed method is evaluated by comparing it with other state-of-the-art methods, showing that it achieves 96.9% accuracy, good performance, and robustness. Hence, it is acceptable to monitor and detect suspicious fall events.

Keywords: Convolutional neural network (CNN); Bi-directional long short term memory (Bi-directional LSTM); you only look once v4 (YOLO-V4); fall detection; computer vision

1 Introduction

The fall event is a familiar suspicious event and a possible hazard to the elderly living in an indoor environment. The main reasons elders are collapsing in physical bodily function could be due to the risk of body imbalance, slow sensory responses, and the risk of being in crowded places such as squares and sports grounds are significant inducements of stampedes [1]. Generally, suspicious falls give rise to injury-related death, immobility, injury, and poor fitness issues in the elderly and some patients. The people's safety is also a significant concern since they stay alone. Manual monitoring of people 24*7 is complex [2]; hence, automatic monitoring and detection of suspicious fall events are essential to security in indoor and outdoor environments.

Over recent years, various methods based on advanced devices have been suggested for suspicious fall event detection. Fast-developing wearable devices that rely on sensors, like tilt sensors, accelerometers, interface pressure sensors, and gyroscopes were extensively adopted in previous research [3]. Previous methods attained high performance in detecting suspicious fall events for the elderly. However, it has been seen that prolonged usage of these devices is unsuitable compared to vision-based devices that can be installed at a specific place and can be used for different people in particular areas. Various vision-based sensors such as RGB cameras, depth sensors, and infrared sensors are utilized for suspicious fall event detection [4]. User-friendly surveillance systems like Red-Green-Blue Color model (RGB) cameras are the most accessible vision-based devices due to the easy installation setup and are cost-friendly. Vision-based devices have been efficiently implemented on popular open-source databases [5]. The objective is to explore suspicious fall event detection with a dynamic background in an indoor environment rather than the static background, which is incompetent to fulfill crucial demands of suspicious fall alarms.

In this paper, we propose a suspicious fall event detection method that can give particular temporal and spatial locations of suspicious fall event detection in an indoor environment while performing daily chores. In our method; an efficient people detector YOLO V4 [6] is first investigated to detect the person and then for multi-object tracking, a method based on deep learning to obtain the trajectories of a person moving from one location to another; afterward, CNN is employed for feature extraction. Subsequently, a visual attention-guided bi-directional LSTM model is proposed to detect suspicious fall events detection. The visual attention-guided model focuses on essential areas of suspicious fall events. The bi-directional LSTM blends the forward within backward time information to anticipate the order consequences of persistent groupings.

The rest of the paper is arranged as follows. Section 2 discusses an outline of the recent works on suspicious fall event detections. Section 3 demonstrates the proposed suspicious fall event detection method, and Section 4 analyzes the experimental results, evaluation of the proposed method, and performance comparison with other existing state-of-the-art methods. Finally, Section 5 concludes the paper with the future exploration of research directions toward suspicious fall event detection.

2 Related Works and Contributions

In the recent decade, suspicious fall event detection methods have been divided into three major classifications: Wearable sensor-based methods, Ambience sensor-based methods, and Computer vision-based methods. This paper focuses on reliable, effective, and most potential deep learning methodologies for computer vision-based methods.

A wearable sensor-based system determines a person's body position information through a wearable sensor device. Most researchers studied accelerometers with different algorithms. The acceleration sensor collects the acceleration of multiple axes, such as gravity values, movement information, and body position information. Primary wearable sensors are cheaper and easier to operate than ambience sensors but highly invasive and uncomfortable, which is a significant drawback [7]. Ambience sensor-based systems place the sensors in an indoor environment and record pressure, radar, audio, vibration, etc. Its primary focus is on wireless techniques to pinpoint changes. It protects privacy without wearing sensor devices on the body; this is relatively costlier [8]. In the revolution of computer vision and image processing methods from 2003–2018, computer vision-based methods became a supreme method since computer systems are cheap, robust, and highly precise. Further background subtraction and feature classification are included with machine learning methods. Suspicious fall event detection was constructed on wide to the long aspect ratio, human body silhouette, shape deformation, etc. In machine learning-based systems, the features had to be identified by supervision, but in deep learning-based

systems, pattern recognition tasks were carried out without feature extraction methods. Deep learning methods are used for lots of development with supervised learning methods using neural networks, making them more effective in recent years than other methods.

Various researchers have been working towards developing efficient deep learning systems to get accurate suspicious fall event detection in real-time video data monitoring [9]. Supervised deep learning methods can be divided into convolutional neural networks (CNN) [10] and long short-term memory (LSTM) [11].

CNN passes video frame images along a sequence of convolutional layers with filters, a pooling layer, a fully connected layer, and a softmax activation function for dataset training and testing. Several researchers categorized CNN for suspicious fall event detection in one dimensional-CNN (1D-CNN), 2D-CNN, 3D-CNN, and Feedback optical flow CNN (FOF-CNN) in previous research. Adhikari et al. [12] proposed high accuracy in video surveillance RGB-depth camera videos to recognize suspicious fall events using CNN for single-person suspicious fall event detection. The system's sensitivity is 99% in the lying position and poor in other positions. The overall accuracy proposed by the system is 74%. Li et al. [13] presented CNN to learn human shape deformation features. Ten-fold cross-validation was executed for training and testing purposes on the dataset. It attains 100% sensitivity and 99.98% accuracy. The performance of the system was not measured in a real-time environment. Yu et al. [14] applied CNN to pre-process extracted human body silhouettes from the background subtraction method. It provides better performance than other machine learning-based systems used previously. Its accuracy observed was 96.88%. Shen et al. [15] detected human key points through the deep-cut neural network model, and these points are relocated into the deep neural network. It shows an accuracy of 98.05%. In 2019 [16], fall events were detected by developing 3D-CNN. It captures both two spatial and temporal information from video frames by Kinect depth camera to achieve 97.58% accuracy with real-life data. Li et al. [17] apply pre-trained data to 3D-CNN to train the model based on Spatio-temporal patterns. It achieved total accuracy with the custom dataset. Hwang et al. [18] proposed 3D-CNN with data augmentation methods and achieved an accuracy of nearly 92.4% to 96.9%. Kasturi et al. [19] suggested a system from a Kinect camera by using video information that forms a stacked cube, and that cube is taken as input to the 3D-CNN. It achieves total accuracy for both the validation and training phases. In 2017 [20], developed an Internet of Things (IoT) system that used 3D-CNN and features a Feedback mechanism scheme for Feedback on safe movement to obtain increased accuracy in less computation.

LSTM is a widely used recurrent structure in sequence modeling. It uses gates to control information flow in recurrent computation which can store and release long-term memories using gates. In previous research, various researchers used LSTM along with CNN, RNN (recurrent neural network), and RCN (recurrent convolutional networks). Lu et al. [21] developed LSTM and 3D-CNN with automatic feature extortion from kinetic data. It solves previous methods' segmentation, occlusion, illumination, and variation problems. The sensitivity achieved was 98.13%, specificity was 99.91%, and accuracy was 99.84%, but the system had not used any real-time data. In 2018, [22] proposed people action classification by 2D skeletons extraction from RGB camera and applied to CNN and then RNN with LSTM. System accuracy of 88.9% was achieved by this method. Abobakr et al. [23] presented a suspicious fall event system based on CNN and RNN. Convolutional LSTM with ResNet is used for feature extraction and attained an average accuracy of 98%. In 2017, [24] analyzed three-dimensional skeleton joints from depth camera input data by employing RNN and LSTM to determine suspicious fall events. Ge et al. [25] introduced Co-Saliency-Enhanced RCN architecture followed by an RNN with LSTM to retrieve results. Experimental results showed 98.96% accuracy. Sultana et al. [26] presented a model in which a 2D Convolutional Neural Network (2DCNN) method was used for extracting the features from the video. To find the temporal dependency of the human movement Gated Recurrent Unit (GRU) network is used, and a sigmoid classifier detected suspicious fall events. The experimental result

shows an accuracy of 99%. In 2022, [27] Proposed VEFED-DL (vision-based elderly fall event detection using deep learning) model to capture the RGB color images from the digital video camera that involves different stages of operations like pre-processing, feature extraction, classification, and parameter optimization. It extracted spatial features, fed into the gated recurrent unit (GRU) to extract the temporal dependencies of the human movements, and through binary classification, detected suspicious fall events with an average accuracy of 99.98%. In 2022, [28] proposed a pre-trained Tensor Flow Lite CNN real-time method to detect key points of a person. LSTM analyzes these key point sequences along with the best hyper-parameter combination. The accuracy of the system was observed 91% by this method.

In the initial studies, suspicious fall event detection systems are investigated based on deep learning methods. Most of the authors have used CNN for the development of suspicious fall event detection. Among all dimensions of CNN, 3D-CNN provides better performance. LSTM contributes to enhancing the performance of suspicious fall event detection systems. Storage and computational power are no longer needed due to the low cost of computing devices in a real-time environment. In comparing the wearable sensor-based and vision-based methods, vision-based methods provide better results than sensor-based methods. Real-time monitoring is also an essential factor. In the suspicious fall event detection method, contacting emergency services such as an ambulance or sending online messages to a mentor or caretaker is mandatory. Systems have to connect for suspicious fall event rescue operations and emergency healthcare services, which were rarely developed by systems. All these to overcome in the proposed work, CNN, along with Bi-directional LSTM deep learning methods, is used to detect suspicious fall event detection for enhanced performance with real-time video data monitoring as input. Once a suspicious fall event is detected, then a rescue operation is provided immediately.

3 Proposed Method for Detection of Suspicious Fall Event Based on Bi-directional LSTM

For the indoor environment, we proposed a bi-directional LSTM for suspicious fall event detection, as shown in Fig. 1, which is divided into three significant chunks: Detection and Tracking of people, bi-directional LSTM layer, and a visual attention-guided method. We initially detected the people in the video frames using YOLO V4. Since the movement of people consistently appears in the sequence of videos, it is essential to track the person. Traditional tracking methods failed to track the absolute trajectories based on distance matching. Hence, we used the deep-sort tracking technique [29] to accomplish the tacking job. From each tracked bounding box of people extracting valuable features is the target. The extractions of human body feature points from the human body contour are valuable features. Valuable features are measured through the contour curves of the body, which are considered connections of a series of line segments. Then, apply the resulting output of the rear convolutional layer of Visual Geometry Group-16 (VGG-16) [30] and redirect all valuable features of every trajectory to the visual attention-guided Bi-directional LSTM model for concluding the suspicious fall event detection.

3.1 Detection of Fall Suspicious Event

People detection is a beneficial and key module in detecting suspicious fall events. Traditional methods for detecting suspicious fall events are delicate towards the shadows, light, noisy backgrounds, and moving objects. We used the detection method YOLO V4 [6]. We saw that it outperforms several one-stage object detector algorithms such as single-shot detector (SSD) [31], YOLO [V1–V3] [32–34], and two-stage object detection algorithms such as Mask Region-based Convolutional Neural Networks (R-CNN) [35], Faster R-CNN [36], Fast R-CNN [37] and R-CNN [38].

YOLO V4 extends the technology of YOLO V3 [34] by adding CSPDarknet53 to enhance the learning capability of CNN. Spatial pyramid pooling is added to increase the receptive field and filter the essential context features. YOLO V4 is replacing Feature pyramid networks (FPN) with PANet (path aggregation

network) for parameter aggregation. YOLO V4 can obtain better accuracy and perform fast by increasing average precision and frames per second in real-time.

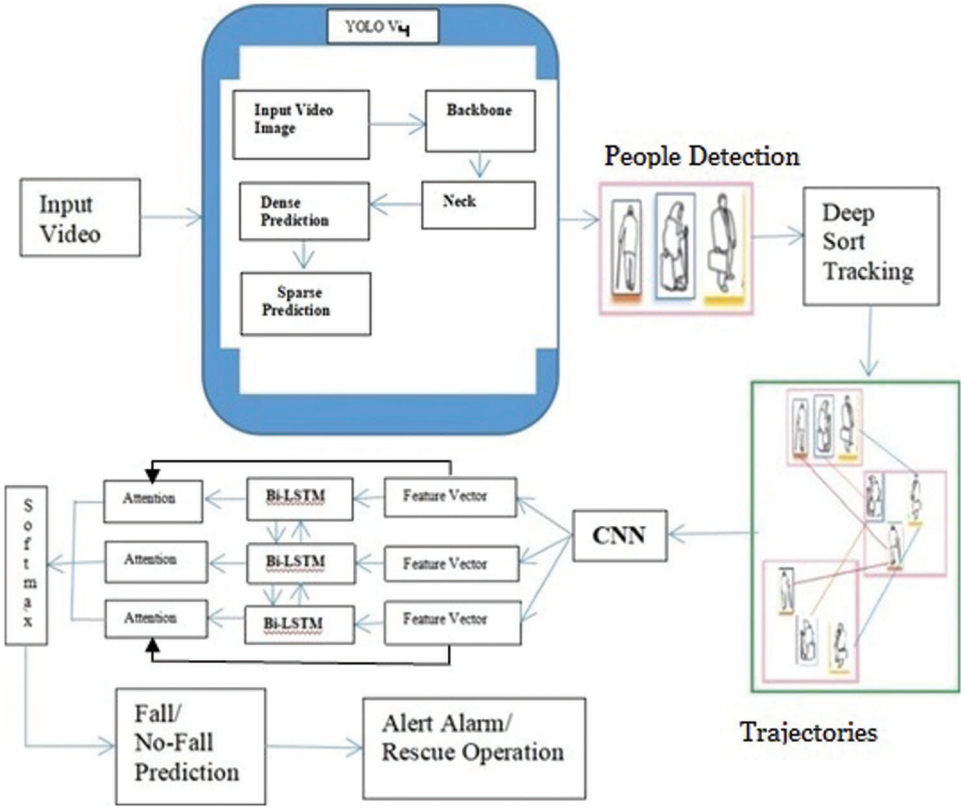

Figure 1: Bi-directional LSTM with visual attention guided architecture for detecting suspicious fall event

3.2 People Tracking

People tracking require tracing people across all sequences of frames in a video. Some common interruptions in real-time tracking are the presence of occlusion, variations in viewpoints, non-stationary cameras, and unrelated training data. Kalman filter and deep learning have shown outstanding results in solving these common interruptions. The deep sort technique uses a distance metric based on the object's appearance and results best among other techniques. The goal of implementing SORT tracking is simple online real-time tracking. Still, due to occlusion, the trajectory of the identical person is disrupted, and a novel trajectory is formed. To avoid complications, the appearance information of bounding boxes is appended, which improves the effect of tracking by matching the preceding trajectory after lengthy occlusion or missed detection. In the Deep-sort technique [29], the Kalman filter is a significant component used to predict a good fit for the bounding box and follow the location of the trajectory. It contains eight-dimensional state information and parameters to track trajectories and delete the trajectories for objects that successfully left the scene, which results in the generation of new bounding boxes. To associate the new bounding box detection with a new prediction, it is required to associate the trajectory with the incoming bounding box by using a distance metric as squared Mahalanobis distance to quantify the association, also to incorporate the uncertainties from the Kalman filter and an efficient and optimized method as Hungarian method to associate data for solving the assignment problem in polynomial time. Then, the appearance feature vector is evaluated by calculating the appearance information of each

bounding box with the help of two convolutional layers followed by a max pool and six residual blocks to generate a single feature vector that is an appearance descriptor that keeps the recent 100 frames in track. The smallest cosine distance is calculated between the appearance feature vectors to resolve the problem of occlusion. The final updated distance is given by: $D = Lambda * DK + (1 - Lambda) * DA$. Where Mahalanobis distance is represented by DK, Cosine distance among the appearance feature vectors is DA, Lambda (Least squares AMBiguity Decorrelation Adjustment) is the factor for weighting; this method is for efficient computation of Kalman filter [29], and D is the final updated distance. The final updated distance for decision information is calculated by combining both distances; Mahalanobis distance is effective for short-term tracking as it gives all possible object locations based on the motion information and the Cosine distance uses appearance information to achieve accurate matching when the object is occluded for a lengthy time. Due to the presence of dense target objects and to make trajectories complete, linear interpolation is used.

3.3 Bi-directional LSTM Layer

Feature vectors are extracted from the convolutional neural network layer. This feature vector gives the output of matrix 15 × 4096 by the full connection layer. (Output Matrix = No. of video frames × No. of filters in a fully connected layer) and transfer the same matrix as input to the bi-directional LSTM. Bi-directional LSTM extends the LSTM by improving the performance by adding one more LSTM. Bi-directional LSTM has one LSTM that takes input in a forward direction, termed forward LSTM, and the second LSTM takes input in a backward direction, termed backward LSTM. Bi-directional learning is much faster and effectively increases the availability of information to the network. In the LSTM [39], forgotten information is determined and retained through the memory controller, which is executed with three gate structures: input gate, forget gate, and output gate.

The working procedure is demonstrated where W_f, W_i, W_g, and W_o represents the preceding feature vector output, U_f, U_i, U_g, and U_o are the current feature vector input across the weight of all control gate, and b_f, b_i, b_g, and b_o are the bias terms passing across the control gate. After passing across the forget gate, the eliminated information is calculated from F_t = sigmoid ($W_t h_{t-1} + u_t x_t + b_t$). To compute the value of state update rate i_t and the state update vector g_t by outputting the door, expressions are evaluated as I_t = sigmoid ($W_i h_{t-1} + u_i x_t + b_i$) and g_t = tanh ($W_g h_{t-1} + u_g x_t + b_g$). Input gate state update rate i_t, forget gate f_t, and the state update vector g_t are provided to compute the update value c_t of c_{t-1} in the internal structure of LSTM by using $C_t = f_t c_{t-1} + i_t g_t$. To determine the specific part of the unit state that will be output across the output gate, expressions are evaluated as P_t = sigmoid ($W_o h_{t-1} + u_o x_t + b_o$) and $h_t = p_t$*tanh(c_t).β. From Fig. 2, the unit structure of the LSTM cell internal architecture can be seen, and the description of various parameters of the LSTM network model is given in Table 1.

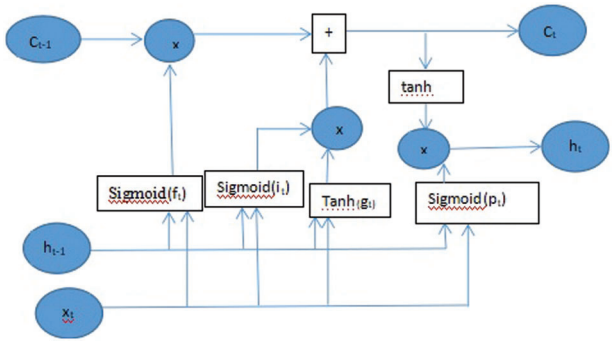

Figure 2: The unit structure of LSTM cell internal architecture

Table 1: Description of various LSTM parameters

Input & output parameters	Description of parameters
C_{t-1}	Previous unit memory.
C_t	Current unit memory.
x_t	At time t, Input to LSTM.
f_t	Determines specifically which part of the old information should be removed.
i_t	Determining the new information should be eliminated or updated.
g_t	Provides weights to the values travel by; importance determining level ranges from −1 to 1.
p_t	Decides output for specific parts of the cell state.
h_{t-1}	Previous LSTM unit output.
h_t	Current network output.

Traditional LSTM network ignores backward-direction learning. It learns only in the forward direction. Despite that, the Bi-directional learns in both directions as the input of the current moment relies upon the preceding video frame and the consequent video frame. So, two LSTM units in different direction combinations observe the temporal information before and after the video frames at current time t. At time t, the unit structure of the Bi-directional LSTM network model [40] is shown in Fig. 3, and the bi-directional LSTM network model's parameters are described in Table 2.

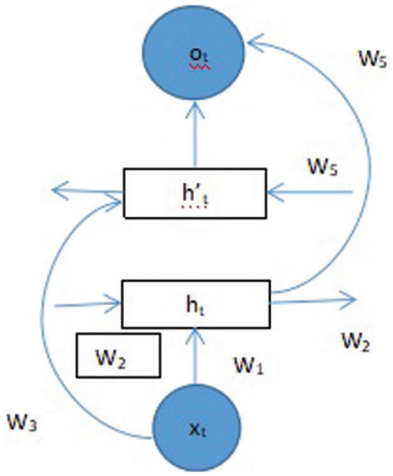

Figure 3: At time t, the unit structure of the Bi-directional LSTM network model

The working procedure is demonstrated by bias in bi-directional LSTM network at time t are represented as $b_t^{(1)}$, $b_t^{(2)}$, $b_t^{(3)}$, and $b_t^{(4)}$ corresponding for both unit directions as well for both directions output. Two-layer counts for bi-directional LSTM and recurrent systems are used. Feature vectors as output from the VGG layer at a particular time deal with both unit directions LSTM results as o'_t and o''_t. VGG layer of 1 × 4096 is used to extract the deep features of the video frame at time t. The feature sequences are considered for full video frame as (…, x_{t-1}, x_t, x_{t+1},…) as given with h_t = sigmoid ($w_1x_t + w_2h_{t-1} + b_t^{(1)}$)

and h'_t = sigmoid ($w_3x_t + w_5h'_{t+1} + b_t^{(2)}$) and the corresponding output is extracted at that particular time by o'_t = tanh ($w_4h_t + b_t^{(3)}$) and o''_t = tanh ($w_6h'_t + b_t^{(4)}$). Average of both output vectors at the parallel time as output feature vector o_t is evaluated by $o_t = (o'_t + o''_t)/2$. The same vector is provided as the input to the visual attention layer to learn the network weight.

Table 2: Description of parameters of the bi-directional LSTM

Parameters	Description of parameters
$W_i \epsilon W_1,\ldots\ldots, w_6$	Weight across one unit layer to another layer.
x_t	VGG layer derived feature vectors from the video frame.
h	Input feature sequences in of LSTM network model in the forward direction.
h'	Input feature sequences in of LSTM network model in the backward direction.
o_t	Output followed by the feature vector passed across the bi-directional LSTM network model.

3.4 Visual Attention Guided Bi-directional LSTM

Visual attention-guided mechanism [41] is essential both spatially and temporally. It points to a few significant features by estimating the weight of the feature vector output at various time steps from the bi-directional LSTM network model. It increases the performance of the whole network model. It is akin to [42] solving the prediction problem of the tracked bounding boxes. It is essential for spatial information as well as for temporal information. It is utilized to record important regions, then preserves the temporal memory and incorporates attention. The visual attention-guided bi-directional LSTM mechanism model is shown in Fig. 4.

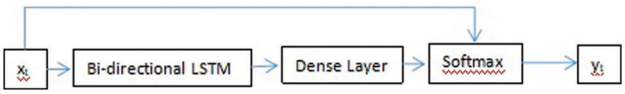

Figure 4: The visual attention-guided bi-directional LSTM mechanism model

Region feature vector at each time-step t is weighted by the visual attention-guided mechanism. Output at time t is evaluated as $Y_t = \sum_{i=1}^{t} \beta_t x_t$. β_t is a softmax activation function over x_t locations and determined as β_t = exp $(w_t o_t)/\sum_t \exp(w_i o_i)$ where t ∈ 1,….,15, and w_i is the Weight mapping to the i^{th} element of the location softmax activation function.

3.5 Automatic Alert Alarm for Rescue Operation

The video data is captured continuously, but the recording is done only for two consecutive spans of one hour. If in the second span of one-hour any suspicious fall event is not reported, the recording of the first span of one hour is deleted, and the space is available for the third span of one hour. This process continues till a suspicious fall event is detected. As trajectories are marked, people are detected, and the lost state of a person from frames is not required to be stored. If during a span, a suspicious fall event is detected, then the system will not delete the current span or the previous span of one hour for further analysis. This reduces the storage requirement without losing information about suspicious fall events.

In case of suspicious fall event detection, an alarm is generated to provide rescue services to people. The word suspicious fall is defined as an event that results in a person coming to rest inadvertently on the ground or at another lower level, and it is identified when the body does not move for one minute. The Joint suspicious fall event defines three types of falls named as forward fall, backward fall, and side fall. To finalize the fall event in suspicious fall event detection, the center of gravity of change rate and aspect ratio of falling person extracted from the bounding boxes, angle of ellipse fitting, and projection histogram were used in the feature vector. There should be no occlusion in the procedure of falling. The whole frame is used as input, and VGG feature vectors are effectively used to detect suspicious fall events, and elimination of mutual interference improves the suspicious fall event detection. The alert is sent to the guardian, caretaker, or mentor. The alert message is in about five seconds [43]. An automatic alert alarm allows the user to call for help without pressing the call button. These systems automatically generate an alert message if the user suffers a fall.

4 Experimental Results

4.1 Dataset and Implement

Two standard datasets were used to develop the comparative study through the experiments. The most popular open source databases, UR fall detection dataset [44] and Multiple Cameras fall datasets [45], are selected for the experiments. Only standard datasets were utilized for the authentication of the databases. Dataset is essential for the training of the model, verification purposes, and performance improvement.

UR Fall Detection Dataset: 30 suspicious fall events and 40 non-fall events mean daily life activity posture videos were shuffled and collected. This dataset has a total of 70 sequences. Suspicious fall events are recorded with two Microsoft Kinect cameras, and non-fall events are recorded with one camera with RGB video data. 70% of data is used for training purposes, the remaining 30% is used for testing purposes, and the evaluation metric precision, recall, F-score, and accuracy parameters have been evaluated.

Multiple Cameras Fall Dataset: 22 suspicious fall events and two non-fall events mean daily life activity videos with multi-view synchronization as 3D scene videos were shuffled, collected, and analyzed. Dataset has a total of 24 scenarios recorded with eight video cameras, and for evaluation metric sensitivity, specificity, and accuracy parameters have been evaluated.

4.2 Metrics

In the dataset, every video consists of frames, and each frame is provided with a label based on the activity performed in that frame. Thus, the multi-class labels are converted to the binary class label of "Fall and Non-Fall" labels. A suspicious event is a tiny subset of an event. As it's a tiny subset, the sample classes are very few, resulting in the class imbalance dataset. Activities of daily living play a role in regular events, while the falling of a person plays the role of a suspicious event. Thus, detecting suspicious fall events is a class imbalance classification problem. The imbalance classification problem deals with performance metrics that focus on the minority classes. Here, in this case, a suspicious fall event is considered the minority class or the tiny subset of the regular activity of daily living.

The Threshold metric is used for evaluating suspicious fall events for imbalanced classification. It quantifies the classification prediction errors and outlines the ratios. Few threshold metrics are used, such as Sensitivity, Specificity, F-score, and Accuracy.

Recall/Sensitivity: The number of correct predicted fall suspicious events to the number of all observed fall suspicious events.

Specificity: The number of correct predicted Non-Fall regular events to the number of all observed Non-Fall regular events.

Precision: The number of correct predicted fall suspicious events to the number of all predicted fall suspicious events.

F-Score: The harmonic mean of precision and sensitivity.

Accuracy: Sum of the number of correct predicted fall suspicious events and the correct predicted Non-Fall events to the sum of all four possible combinations of the confusion matrix.

Confusion Matrix: It defines the observed outcomes as True and False and predicted outcomes as Positive and Negative. It has four possible combinations, as shown in Table 3.

Table 3: Confusion matrix

Observed Predicted	Fall suspicious event (1)	Non-Fall regular event (0)
Fall suspicious event (1)	TP (True Positive)	FP (False Positive)
Non-Fall regular event (0)	FN (False Negative)	TN (True Negative)

4.3 Performance Comparison of People Detection and Tracking

People detection is a prime procedure in suspicious fall event detection. To a large extent, it carries through the effectiveness of feature extraction. To demonstrate the advantages of the YOLO V4 method, a comparison is performed based on the three preferred state-of-the-art methods as the baseline for comparison. The performance is shown in Table 4 on the UR Fall Detection Dataset.

Table 4: Performance comparison of methods proposed on UR Fall detection dataset

Metrics Method	Precision (Unit-%)	Recall (Unit-%)	F-score (Unit-%)	Accuracy (Unit-%)
Frame difference	85.1	72.6	75.5	–
GMM	84.5	81.0	82.7	–
GMG	86.0	71.2	81.3	–
Type-2 Fuzzy gaussians mixture model	84.8	79.3	82.1	–
Type-2 Fuzzy gaussians mixture model with Markov random field	85.9	78.6	81.3	–
Fuzzy adaptive Self-Organizing Map network	80.8	76.1	79.6	–
Practical ReProCS	84.6	80.1	82.7	–
Incremental principal component pursuit	86.0	80.9	83.5	–
Faster R-CNN	96.32	99.80	–	96.27
Mask R-CNN	89.7	81.3	85.2	–
Proposed method (YOLO V4)	89.9	81.5	85.4	95.7

1. Traditional Background Subtraction [35]: There are eight major methods that are effective in the procedure-Frame difference [46], Gaussian mixture model (GMM) [47], Geometric multi-grid (GMG) [48], Type-2 Fuzzy Gaussians Mixture Model [49], Type-2 Fuzzy Gaussians Mixture Model with Markov Random Field [50], Fuzzy adaptive self-organizing map networks [51], Practical ReProCS [52] and

Incremental principal component pursuit [53]. All these works are traditionally well but majorly affected by the object movement and ground shadows, which results in poor performance in the recall, F-score, and precision matrix.

2. Faster R-CNN: It demands cooperation with tracking methods to extract a single-person bounding box. It can extract accurate location information and cannot obtain contour information from the foreground object [54].

3. Mask R-CNN: It demonstrates improvement over the background subtraction methods. It shows better performance, greater robustness, and ample contour information of the human body, but it is a slow method compared to YOLO V3 [55].

4.4 Performance Comparison with the State-of-the-Art Methods

To impartially authenticate the effectiveness of the proposed work with YOLO V4 and the visual attention-guided Bi-directional LSTM, a comparison is performed based on the performance of the proposed method with three-state-of-the-art methods based on deep learning using the standard Dataset of UR Fall Detection Dataset shown in Table 5 and by using Multiple Cameras Fall Dataset is shown in Table 6. The three methods which are compared for performance are given below:

Table 5: Performance comparison of methods proposed on UR fall detection dataset

Metrics Method	Precision (Unit-%)	Recall (Unit-%)	F-score (Unit-%)	Accuracy (Unit-%)
CNN-FD	–	100	–	95.0
CNN-LSTM-FD	94.8	91.4	93.1	–
Mask-RCNN-Bi-directional LSTM-FD	100	91.8	94.8	96.7
Proposed method (with attention-guided LSTM)	100	91.7	94.6	96.6
Proposed method (with attention-guided Bi-LSTM)	100	91.9	95.8	96.9

Table 6: Performance comparison of methods proposed on multiple cameras fall dataset

Metrics Method	Sensitivity (Unit-%)	Specificity (Unit-%)
CNN-LSTM-FD	91.6	93.5
Proposed method (with attention-guided LSTM)	91.8	94.8
Proposed method (with attention-guided Bi-LSTM)	92.0	96.0

1. Convolutional Neural Network Fall Detection (CNN-FD) [56]: As input, optical flow images are taken and applied to two-dimensional convolutional neural networks. It makes decisions based on the incorporation of optical information via motion information.

2. Convolutional Neural Network–LSTM-Fall Detection (CNN-LSTM-FD) [35]: It uses the YOLO V3 framework for obtaining the human body bounding box; foreground features are extracted through CNN along with attention-guided forward direction, single-dimensional LSTM network for fall detection.

3. Mask-RCNN-Bi-directional LSTM-Fall Detection (Mask-RCNN-Bi-directional LSTM-FD) [55]: For detection of the people in frames, Mask-RCNN is applied. After people detection from individual binary

images of the human body contour, essential features are extracted. The endmost VGG-16 convolutional layer output and essential features of individual binary images are fed to visual attention-guided bi-directional LSTM for fall detection.

4.5 Evaluation

The proposed method achieves better performance in person detection and tracking procedures by taking the whole frame as input. It provides VGG features effectively after tracking each person's trajectory. An independent image sequence is generated by eliminating the mutual interference, which results in improved performance. 5-fold cross-validation is applied to all the methods on the training data for impartial comparison with other state-of-the-art methods. With the effective data analysis results along with the support of the deep neural network, it has been seen that deep learning methods do not require any hand-craft features. The concluding result mainly depends on the extracted human body from the bounding box. All results of experiments demonstrate that the proposed method is stable, has an anti-noise performance, and the visual attention-guided bi-directional LSTM improves performance over visual attention-guided forward direction LSTM.

5 Conclusion

A suspicious fall event based on the YOLO V4 with visual attention-guided bi-directional LSTM is proposed in the paper. In our method, YOLO V4 is illustrated in the people detection method, and deep-sort tracking is adopted to track a person for the tracking method. The VGG16 CNN model is adopted for extracting features of individual trajectories. Moreover, the following Bi-directional LSTM provides a visual attention-guided model in temporal and the other spatial region in forward and backward directions for proper classification. The visual attention-guided model and bi-directional LSTM model improve the performance of suspicious fall event detection. All results of experiments illustrate that the proposed method can effectively perform accurate suspicious fall event detection in video and outperform state-of-the-art methods. In the Future, the proposed work can also be utilized for crowded outdoor environments to protect people's lives.

Acknowledgement: We are thankful to the University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, for continuous support, motivation, and cooperation, which helped us carry out our work successfully.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Z. Rubenstein, "Falls in older people: Epidemiology, risk factors and strategies for prevention," *Age and Ageing*, vol. 35, no. 2, pp. ii37–ii41, 2006.

[2] M. Agrawal and S. Agrawal, "Suspicious event detection in real-time video surveillance system," in *Social Networking and Computational Intelligence, Int. Conf., SCI-2k18. Proc.: Lecture Notes in Networks and Systems (LNNS 100)*, Bhopal, India, pp. 509–516, 2020.

[3] R. Rucco, A. Sorriso, M. Liparoti, G. Ferraioli, P. Sorrentino *et al.,* "Type and location of wearable sensors for monitoring falls during static and dynamic tasks in healthy elderly: A review," *Sensors*, vol. 18, no. 5, pp. 1613–1633, 2018.

[4] J. Han, L. Shao, D. Xu and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[5] L. Maddalena and A. Petrosino, "Background subtraction for moving object detection in RGBD data: A survey," *Journal of Imaging*, vol. 4, no. 5, pp. 71–98, 2018.

[6] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *Computer Vision and Pattern Recognition*, pp. 1–17, 2020. https://doi.org/10.48550/arXiv.2004.10934

[7] T. Xu, Y. Zhou and J. Zhu, "New advances and challenges of fall detection systems: A survey," *Applied Sciences*, vol. 8, no. 3, pp. 418–429, 2018.

[8] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang *et al.,* "RT-fall: A real-time and contactless fall detection system with commodity WiFi devices," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 511–526, 2017.

[9] M. M. Islam, O. Tayan, M. R. Islam, M. S. Islam, S. Nooruddin *et al.,* "Deep learning based systems developed for fall detection: A review," *IEEE Access*, vol. 8, pp. 166117–166137, 2020.

[10] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. A. Dujaili, Y. Duan *et al.,* "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 53–127, 2021.

[11] G. V. Houdt, C. Mosquera and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, pp. 5929–5955, 2020.

[12] K. Adhikari, H. Bouchachia and H. N. Charif, "Activity recognition for indoor fall detection using convolutional neural network," in *IEEE Xplore, 15th IAPR Int. Conf. on Machine Vision Applications (MVA 2017)*, Nagoya, Japan, pp. 81–84, 2017.

[13] X. Li, T. Pang, W. Liu and T. Wang, "Fall detection for elderly person care using convolutional neural networks," in *IEEE Engineering in Medicine and Biology, 10th Int. Congress on Image and Signal Processing-BioMedical Engineering and Informatics (CISP-BMEI 2017)*, pp. 1–6, 2018.

[14] M. Yu, L. Gong and S. Kollias, "Computer vision based fall detection by a convolutional neural network," in *ACM, 19th ACM Int. Conf. on Multimodal Interaction (ICMI 2017)*, Glasgow, UK, pp. 416–420, 2017.

[15] L. Shen, Q. Zhang, G. Cao and H. Xu, "Fall detection system based on deep learning and image processing in cloud environment," *Advances in Intelligent Systems and Computing, 12th Int. Conf. on Complex, Intelligent, and Software Intensive Systems (CISIS-2018)*, vol. 772, pp. 590–598, 2019.

[16] M. Rahnemoonfar and H. Alkittawi, "Spatio-temporal convolutional neural network for elderly fall detection in depth video cameras," in *IEEE, IEEE Int. Conf. on Big Data (Big Data)*, Seattle, WA, USA, pp. 2868–2873, 2018.

[17] S. Li, H. Xiong and X. Diao, "Pre-impact fall detection using 3D convolutional neural network," in *IEEE, IEEE 16th Int. Conf. on Rehabilitation Robotics (ICORR 2019)*, Toronto, Canada, pp. 1173–1178, 2019.

[18] S. Hwang, D. Ahn, H. Park and T. Park, "Poster abstract: Maximizing accuracy of fall detection and alert systems based on 3D convolutional neural network," in *IEEE, Second ACM/IEEE Int. Conf. on Internet-of-Things Design and Implementation (IoTDI 2017)*, Pittsburgh, PA USA, pp. 343–344, 2017.

[19] S. Kasturi, A. Filonenko and K. H. Jo, "Human fall recognition using the spatiotemporal 3D CNN," in *The 24th Int. Workshop on Frontiers of Computer Vision (IW-FCV2018)*, Hakodate, Hokkaido, Japan, pp. 1–3, 2019.

[20] Y. Z. Hsieh and Y. L. Jeng, "Development of home intelligent fall detection IoT system based on feedback optical flow convolutional neural network," *IEEE Access*, vol. 6, pp. 6048–6057, 2018.

[21] N. Lu, X. Ren, J. Song and Y. Wu, "Visual guided deep learning scheme for fall detection," in *IEEE Xplore, 13th IEEE Int. Conf. on Automation Science and Engineering (CASE 2017)*, Xi'an, China, pp. 801–806, 2018.

[22] W. N. Lie, A. T. Le and G. H. Lin, "Human fall-down event detection based on 2D skeletons and deep learning approach. Chiang Mai, Thailand, pp. 1–4, 2018.

[23] A. Abobakr, M. Hossny, H. Abdelkader and S. Nahavandi, "RGB-D fall detection via deep residual convolutional LSTM networks," in *IEEE Xplore, Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, Australia, pp. 1–7, 2018.

[24] X. Tao and Z. Yun, "Fall prediction based on biomechanics equilibrium using Kinect," *International Journal of Distributed Sensor Networks*, vol. 13, no. 4, pp. 1–9, 2017.

[25] C. Ge, I. Y. H. Gu and J. Yang, "Co-saliency-enhanced deep recurrent convolutional networks for human fall detection in e-healthcare," in *IEEE Xplore, 40th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, Hawaii, pp. 1572–1575, 2018.

[26] A. Sultana, K. Deb, P. K. Dhar and T. Koshiba, "Classification of indoor human fall events using deep learning," *Entropy*, vol. 23, no. 1, pp. 328–348, 2021.

[27] G. Anitha and S. Baghavathi Priya, "Vision based real time monitoring system for elderly fall event detection using deep learning," *Computer Systems Science & Engineering*, vol. 42, no. 1, pp. 87–103, 2022.

[28] N. Mamchur, N. Shakhovska and M. G. ml, "Person fall detection system based on video stream analysis," in *Procedia Computer Science, Proc. of Int. Workshop on Small and Big Data Approaches in Healthcare (SBDaH-2021)*, Leuven, Belgium, vol. 198, pp. 676–681, 2022.

[29] N. Wojke, A. Bewley and D. Paulus, "Simple online and real time tracking with a deep association metric," in *IEEE Xplore, IEEE Int. Conf. on Image Processing (ICIP 2017)*, Beijing, China, pp. 3645–3649, 2017.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. on Learning Representations (ICLR 2015)*, San Diego, CA, USA, pp. 1–15, 2015.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multiBox detector," in *European Conf. on Computer Vision (ECCV). Proc.: Lecture Notes in Computer Science*, Cham, Switzerland, 9905, pp. 21–37, 2016.

[32] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Xplore, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779–788, 2016.

[33] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Xplore, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6517–6525, 2017.

[34] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *Computer Vision and Pattern Recognition*, pp. 1–6, 2018. https://doi.org/10.48550/arXiv.1804.02767

[35] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song *et al.,* "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognition Letters*, vol. 130, pp. 242–249, 2018.

[36] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *ACM, Proc. of 28th Int. Conf. on Neural Information Processing Systems*, Montreal, Canada, 1, pp. 91–99, 2015.

[37] R. Girshick, "Fast R-CNN," in *IEEE Xplore, Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 1440–1448, 2015.

[38] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Xplore, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580–587, 2014.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] A. Graves, S. Fernández and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Artificial Neural Networks: Formal Models and Their Applications (ICANN 2005)*, Berlin, Germany, 3697, pp. 799–804, 2005.

[41] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd Int. Conf. on Learning Representations (ICLR 2015)*, San Diego, CA, USA, pp. 30–45, 2015.

[42] S. Sharma, R. Kiros and R. Salakhutdinov, "Action recognition using visual attention," in *4th Int. Conf. on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, pp. 1–11, 2016.

[43] J. R. Abbe and C. O'Keeffe, "Continuous video monitoring: Implementation strategies for safe patient care and identified best practices," *Journal of Nursing Care Quality*, vol. 36, no. 2, pp. 137–142, 2020.

[44] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[45] E. Auvinet, C. Rougier, J. Meunier, A. S. Arnaud and J. Rousseau, "Multiple cameras fall data set," Technical Report Number 1350, DIRO-University of Montreal, 2010.

[46] M. Krekovic, P. Ceric, T. Dominko, M. Ilijas, K. Ivancic *et al.,* "A method for real-time detection of human fall from video," in *IEEE Xplore, Proc. of the 35th Int. Convention MIPRO*, Opatija, Croatia, pp. 1709–1712, 2012.

[47] A. Poonsri and W. Chiracharit, "Fall detection using gaussian mixture model and principle component analysis," in *IEEE Xplore, 9th Int. Conf. on Information Technology and Electrical Engineering (ICITEE)*, Phuket, Thailand, pp. 1–4, 2017.

[48] A. B. Godbehere, A. Matsukawa and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *IEEE Xplore, Proc. of American Control Conf. (ACC)*, Montreal, QC, Canada, pp. 4305–4312, 2012.

[49] F. El Baf, T. Bouwmans and B. Vachon, "Type-2 fuzzy mixture of Gaussians model: Application to background modeling," in *Advances in Visual Computing, 4th Int. Symp. on Visual Computing (ISVC)*, Las Vegas, NV, USA, pp. 772–781, 2008.

[50] Z. Zhao, T. Bouwmans, X. Zhang and Y. Fang, "A fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Communications in Computers and Information Science (CCIS), Int. Conf. on Multimedia and Signal Processing (CMSP)*, Berlin, Heidelberg, vol. 346, pp. 177–185, 2012.

[51] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[52] H. Guo, C. Qiu and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4284–4297, 2014.

[53] P. Rodriguez and G. Chau, "Panning and jitter invariant incremental principal component pursuit for video background modeling," in *IEEE Xplore, Proc. of IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, Venice, Italy, pp. 1844–1852, 2017.

[54] W. Min, H. Cui, H. Rao, Z. Li and L. Yao, "Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics," *IEEE Access*, vol. 6, pp. 9324–9335, 2018.

[55] Y. Chen, W. Li, L. Wang, J. Hu and M. Ye, "Vision-based fall event detection in complex background using attention guided Bi-directional LSTM," *IEEE Access*, vol. 8, pp. 161337–161348, 2020.

[56] A. N. Marcos, G. Azkune and I. A. Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–16, 2017.